BUSINESS STATISTICS

BY

GEORGE R. DAVIES

PROFESSOR OF BUSINESS STATISTICS
COLLEGE OF COMMERCE, UNIVERSITY OF IOWA

AND

DALE YODER

PROFESSOR OF ECONOMICS AND INDUSTRIAL RELATIONS
SCHOOL OF BUSINESS ADMINISTRATION
UNIVERSITY OF MINNESOTA

SECOND EDITION

NEW YORK

JOHN WILEY & SONS, INC.

LONDON: CHAPMAN & HALL, LIMITED

PREFACE TO SECOND EDITION

The present edition represents a number of modifications of the materials presented in the first edition, many of them suggested by instructors as a result of their experience with the earlier volume. Most notable, perhaps, is the rather comprehensive reorganization of the materials, their division into several new chapters. The new arrangement more closely approximates the traditional presentation, and it has facilitated an elaboration of certain phases, notably graphic representation, analysis of seasonal variations, and variance analysis.

In the first edition, the authors ventured to depart somewhat from the traditional content of courses in business statistics in order to take account of "small-sample" theory, as illustrated in measures of significance and analysis of variance. This innovation appeared justified by the fact that, in recent years, small-sample theory has been greatly improved, and its application to the data of business and economics is highly desirable. Hence, although the established methods of large-sample analysis were generally followed in the earlier edition, some attention was directed to these recent developments.

Small-sample theory as applied particularly to the determination of significance has now definitely established itself in business practice. In the present edition, therefore, the subject is further elaborated and is introduced in elementary form early in the course. Its application is extended in so far as it accords with advanced business practice. Every effort has been made to simplify the analysis so as to make it readily understandable by students without specialized mathematical training.

It should be remembered, however, that the problem of determining reliability is not limited to small-sample theory, but is very much broader. Whether statistics like averages and correlation coefficients are reliable or not depends primarily upon the nature of the environment from which they are drawn. Business situations change from boom to depression and from peacetime

v

to wartime activities, so that normals computed and conclusions established during one period may not necessarily hold in another. There is no mathematical formula for determining reliability in this sense of the term. A broad understanding of the whole business and social situation is essential to sound judgment. However, in fairly stable situations, as for example in agricultural experimentation, and to a limited extent in personnel and related problems, measures of statistical reliability based upon the theory of probability are useful.

It is important to emphasize, however, that the theory of mathematical probability is built upon oversimplified experiments in coin tossing or similar expressions of chance, and it makes the assumption of so-called binomial or normal distributions in the field from which the samples are drawn. Practically nowhere in statistical practice, as in biology, sociology, or economics, do we find anything as elementary and simple as the theory assumes. Hence both extreme caution and considerable experience are required in adapting and interpreting measures of statistical probability to the complex data of business. Obviously, conclusions thus based should be considered tentative approximations even at the best, and not absolute or exact as stated. This caution is particularly important in complex fields such as correlation, especially in relation to time series.

GEORGE R. DAVIES
DALE YODER

PREFACE TO FIRST EDITION

It is the purpose of this textbook to present the elementary processes of statistical analysis from the standpoint of business practice with a minimum of mathematical interpretation. Account has been taken, however, of the increasing emphasis upon problems of reliability and significance, particularly as approached in the recent contributions by R. A. Fisher and G. W. Snedecor. Correlation is presented as a development of trend fitting with a view to its predictive applications in rapidly developing fields such as personnel management. Otherwise, the conventional outline of business statistics has been followed. The emphasis is placed upon principles and fields of application, the derivation of formulas and more specialized techniques being relegated to the Appendix, which also contains the more commonly used statistical tables.

Although general statistical methods are universal in their application, there is unquestionable hazard in their widespread use, and intelligent interpretation is essential. The authors, therefore, recognize that what is needed today, at least as much as a knowledge of statistical techniques, is an understanding of their limitations, of the instances when such methodology is inapplicable, and of the caution which should characterize their application.

It is assumed, however, that a knowledge of abstract methodology is necessary before the necessary limitations upon its use can be comprehended. Students must know how statistical techniques are customarily used before they can recognize the limitations and dangers of misuse. For this reason, primary emphasis is placed upon explanation and illustration of method, after which attention is directed to the common errors of misuse. It is hoped that a comprehensive understanding of the limitations of statistical method will be provided by subsequent excursions into the field.

It will be noted that the text has been divided into two parts.

Part I, consisting of Chapters I, II, III, IV, and V, includes the material most commonly encountered in a single semester or a quarter course in business statistics. For continuation courses involving an additional semester, Part II, consisting of Chapters VI, VII, VIII, IX, and X, presents the more complex applications of various types of correlation and of measures of reliability and significance, including a brief treatment of the analysis of variance.

GEORGE R. DAVIES
DALE YODER

CONTENTS

PART I

PART II

BUSINESS STATISTICS

PART I

. CHAPTER I

THE FIELD OF BUSINESS STATISTICS

At the outset, it may be well to consider the nature of statistics and the objectives of study in the field of business and economic statistics.

Statistics, according to Webster's Dictionary, is the science that deals with " the collection and classification of facts on the basis of the relative number or occurrence. . . ." The most significant aspects of this definition are three in number: (1) Statistics is a science, i.e., a search for information and knowledge, a program designed to furnish facts. It is an applied or useful science or art of the same category as engineering or dentistry, and it is based fundamentally on the pure science of mathematics. (2) Its elementary process is the collection and classification of data, which is, after all, a basic technique of all science. (3) Statistics places special emphasis on the quantitative character of the data it classifies. Qualities are, of course, frequently utilized as criteria in classification. But statistics notes particularly the frequency of occurrence, and elements of the study include collection, identification, classification, analysis, presentation, and synthesis of data.

A simple illustration may serve to suggest the elementary features of statistical analysis and synthesis. Suppose, for instance, that it is desired to know the average family consumption of ice cream throughout a given locality and the significant differences in consumption among families receiving

different amounts of income. Any such study would lean heavily upon statistics to attain its objectives. If data were secured from every family, the data thus obtained would have to be analyzed to discover the most important *common* or *general* characteristics of the group. Some sort of average measure of consumption would probably be taken, and possibly such an average would be found for each of several income classes.

Very frequently, in such a study, however, the task of securing information from each individual family is regarded as too expensive and time-consuming. In such cases, resort is had to sampling, another process in which statistical analysis plays an important part and one that will be described in detail in later chapters. Steps must be taken to discover how a truly representative sample can be secured. Perhaps the families in the state will be classified according to such criteria as size; location in village, town, city, or rural districts; amount of income; and others, so that all these classes may be proportionately represented in the sample. This type of preliminary analysis has been given great impetus in recent years by the various polls of public opinion, all of which make use of this sort of sampling. Another related statistical problem would consider the necessary size of the sample to be taken.

Types of problems.—The variety of statistical problems in economics and business is obviously vast. Sometimes, effort is made to differentiate between business problems and economic problems. In general, such a distinction regards problems of business statistics as those reflecting the internal control of business concerns, while economic statistics deal with interbusiness relationships. Such a distinction, however, is difficult to maintain, since no sensible business management can afford to ignore interbusiness relationships or their significance for internal business policies and practices. Practically, therefore, the two terms *business statistics* and *economic statistics* may be regarded as synonymous, at least so far as the techniques they employ and the data with which they deal are concerned.

On the other hand, the field may be divided, for convenience of description, into four principal sections. In one of them,

major attention is given to the use of statistics by management in appraising internal managerial and financial policies. In a second, management uses statistics as a guide to merchandizing its products or to its external policies, those upon which its dealings with competitors and consumers are based. Again, as a third major field, there is the use of statistics by governmental agencies as a basis for various social and political policies having economic implications.

It is desirable, also, that a fourth field should be described, one that is sometimes regarded as the whole field of *economic* statistics, that in which statistical techniques are used to check deductive reasoning in the establishment or disestablishment of economic principles. Thus the generalization usually described as Engel's law, which holds that proportionate shares of family income expended for necessities decline as family income increases, might be subjected to study through statistical analysis of family budgets. Again, studies of the elasticity of demand for various commodities or services may be made to discover evidence supporting or refuting assumptions as to that elasticity.[1]

Managerial problems: production.—Each of these fields deserves brief consideration in an outline of the field. The uses of statistics by management may be broadly classified as those dealing with production and those related to distribution.

In production, modern management finds many opportunities for profitable statistical analysis. Its purchases of raw materials may well be based on analysis of the comparative characteristics of such materials secured from various sources. Some sources may provide better materials, with consequent reduced waste and cheaper processing. The establishment and maintenance of standardized production processes likewise depend on careful statistical analysis. Production standards on individual processes are most satisfactorily set on a basis of careful observation, repeated timing, and averaging. Planning of production requires similar detailed analysis of the timing of

[1] See, however, George J. Stigler, "The Limitations of Statistical Demand Curves," *Journal of the American Statistical Association*, 34 (207), September, 1939, pp. 469–481.

various processes and the interrelationships among them. Transportation, both of raw material and of finished products, may well consider comparative costs as well as such features as promptness, dependability, and convenience. Statistical analysis of each of these features may be necessary to supplement casual judgment.

Throughout this production process, the problems of personnel management play an important role. Here, again, statistical analysis contributes greatly to the perfection of satisfactory policies. Sources of workers may well be compared; selection methods, including questionnaires, tests, and others, may be appraised; types of compensation and their effects on productive efficiency may be contrasted; and sources of friction, complaints, and inefficient performance may be evaluated. All these types of analysis will frequently depend upon statistical techniques. At the same time, consideration must be given to questions involving the extension of credit and the effective use of all financial resources. Here, again, facts must be gathered and classified. Every phase of management, therefore, leans upon the compilation, classification, analysis, and synthesis of pertinent data of a statistical nature.

Managerial problems: distribution.—When one turns to the external aspects of an individual business, the opportunities for profitable statistical analysis are at least as great. Basic to all of them is a measurement of potential markets, the analysis of consumption. Alert managements need to know how much of their product can be sold at various prices. They seek to discover who uses their product and who uses something else which their product might replace. They want to know where their best markets are located, whether in city or country, in what sections of the country, and in what potential markets they are failing to secure the sales they should have. They wish to compare the effectiveness of various types of advertising: the radio, the newspaper, and the magazine. They may seek to appraise various systems of premiums, special offers, and other introductory campaigns.

In some lines, it may be desirable to attempt a measure of the elasticity of demand and thus to determine the significance

of possible price changes. In others, the whole long-time trend of consumption may be of great importance. Thus, in recent years, life-insurance companies have made extensive analyses to determine whether their product has reached its peak in this country, whether life insurance is to continue the growth characteristic of the past sixty years or is to level off at approximately present totals. Similarly, employment stabilization is to a considerable degree dependent upon careful measurement of seasonal and longer-time fluctuations in employment.

So far as distributive aspects include retail and wholesale establishments they introduce numerous additional statistical problems. In retail stores, in addition to personnel problems similar to those already suggested, there are studies of the value of credit extensions and of installment selling, analyses of sales potentialities of various items to determine what space and location shall be given to such items, checking of potential sales volume as a factor in store location, forecasts of sales as a basis for purchasing and maintaining inventories, and numerous other studies. Wholesale units have similar problems. Effects of price policies, price-fixing legislation, price agreements with other distributors or with manufacturers, all require statistical analysis as a basis for appraisal.

Statistical analysis and social policy.—There are, of course, numerous situations in which statistical analysis may be used effectively to provide information essential to sound social policy with respect to current economic problems. Societies utilize such information in order that they may so adjust governmental practices that desired objectives may be attained. In the absence of reliable information or appropriate analysis and interpretation, social policy may actually prevent the attainment of desired goals, or it may achieve only part of what it could accomplish.

Many of the most elementary economic data needed by modern societies are available only as a result of extensive statistical analysis. This is true, for instance, of data on the national income and the distribution of that income among families and individuals or among the various states. It is similarly true of data on exports and imports, the volume of production and con-

sumption, and the amount of saving and investment. Yet it will be obvious that such data are essential to intelligent government in any modern society.

The same situation prevails with respect to employment and unemployment. In the United States, for example, all current monthly data as to numbers unemployed are estimates arrived at by fairly complicated statistical analysis of reports from a relatively small number of firms, industries, and trades. There is no actual counting of all the unemployed except at the ten-year census intervals. Even when a special project to enumerate the unemployed is undertaken, as was done in 1937, the result is adjusted by checking against data more carefully collected for a few sample areas.

. The same dependence upon statistical analysis and interpretation is evident in numerous other connections. Thus, information about changes in money wages and real wages, fluctuations in business activity, changes in rates of interest, securities sales, and many other aspects of finance is available only as a result of the compilation, classification, and analysis of large quantities of data. The number and variety of subjects upon which current data are collected and subjected to analysis may be most readily observed by studying the pages of the *Survey of Current Business* or examining the index of the *Statistical Abstract of the United States*. A wide range of governmental policies leans heavily upon such information, and programs of taxation, relief, education, regulation, and numerous others can be intelligently administered only if current data are reliable, accurate, and properly interpreted.

Statistics and economic theory.—In one of its most important uses, statistical analysis, as has been said, provides a means of checking theoretical conclusions as to various economic relationships and suggesting still other, similar relationships. Thus, the widely described principle of wage determination generally described as the marginal productivity theory has been subjected to critical appraisal and thereby substantiated as an historical reality by the statistical investigations of Professor Paul Douglas.[1] Similarly, the studies of Professor Henry

[1] *The Theory of Wages*, New York, Macmillan Co., 1934.

Schultz have provided factual data to substantiate theoretical deductions as to the historical elasticity of demand for a variety of commodities. These are but pioneering examples of much more extensive analysis that may be made. Their conclusions may, and doubtless will, cast serious doubt on many of the principles arrived at by deduction. Other such principles will be given greater reliability by the statistical evidence derived from analysis of current and historic data.

Contrasts between physical and social sciences.—The striking successes of the physical sciences in the past four centuries have not yet been paralleled by the social sciences. It is true, of course, that there has been some improvement in theoretical thinking with respect to economics and sociology. The modern world unquestionably has a somewhat more accurate theoretical conception of the laws of supply and demand and of investment and capitalization than the ancients had. But, with respect to a generally acceptable understanding of the workings of the modern, world-wide economic system and with respect to the elementary nature of that organization, modern society can demonstrate no great advance over the philosophical hypotheses of ancient Greeks.

A philosopher from historic Athens, if transplanted into the modern economy, would be utterly bewildered at its complex machinery, but he might be quite at home in our conceptions of property, trade, and law. The atomic theory has been so expanded and applied that its founder, Leucippus, could not, without long study, comprehend modern physical, chemical, and electrical laboratories, but Aristotle would have little difficulty in understanding the principles generally described as operative in our economic system. Similarly, the Roman lawyer could, without serious problems, transform his *jus gentium* into the more complex form of modern corporation law. The historic increase in domestic and international trade, the hostilities and wars, the tendency toward dictatorships, would all be familiar features to the thinkers of the ancient and medieval worlds. The social sciences in modern society stand in much the same position as that of the physical sciences three centuries ago. Only a beginning has been made in the analysis of actual facts;

major dependence has been placed on deductions from the realm of abstraction. A few attempts at inductive study have been made, and, as has already been noted, partial index numbers and incomplete summaries have been prepared to provide a crude picture of the larger changes in business. Even this meager beginning of the development of the factual aspect is forcing numerous revisions of social and economic theory.

Problems of depression.—Recently, popular as well as professional interest in the functioning of the modern economic system has been focused on problems associated with depression. The world-wide recession following 1929 occasioned such extensive social losses in terms of idle men and machinery that the question of how industry may be maintained constantly at a reasonably high level of productivity is of utmost practical concern. Various possible causes of business recession, in addition to those involving the rates of saving, have been advanced by exponents of differing theories, including such diverse and frequently contradictory conditions as the excessive or deficient quantity of money in circulation, the presence of restrictions on international trade, the inefficient allocation of capital equipment with respect to future demand, the development of monopoly in both capital and labor with consequent "sticky" prices, the absence of sufficient cooperation in capital and labor to maintain "reasonable" prices, and the discrepancy of the marginal efficiency of capital and the interest rate. At the same time, a host of panacean "cure-all" remedies has appeared and received varying degrees of popular support. Some would "spend our way to recovery"; others would "balance the budget" as a certain way out. Still others propose an "economy of scarcity" or an "economy of abundance," while some have concluded that "economic planning" or a "rubber dollar" or "parity prices" or "guaranteed costs of production" represent the simple but final solution. Advocates have urged that wage rates be raised, that "high wages is good business," that what we need is more "purchasing power." Still others insist that wages must be deflated, that hours must be reduced, that production must be restricted, that "every third row must be

plowed under," that saving must be curtailed, that the trouble arises from the practices of "economic royalists."

Not all these conflicting theories can be checked with the facts in a society as complex as that of the modern world, but approximate checks are possible for some of them, and there is ample justification for the collection of data to this end. However, it is impossible to secure these data and analyze them if only a few investigators undertake the task. A comprehensive, long-time program, and the cooperation of endowed private bureaus and governmental agencies are essential.

The point is that the programs currently suggested and undertaken to expand and maintain production upon the high levels essential to comfort and security in modern society are founded upon little more than speculation and prejudice. This is true in abstract theory as well as in concrete politics. These policies and programs can secure a firmer, more effective foundation only as the processes with which they deal are better understood, as facts and measures replace conjectural hypotheses, beliefs, folklore, and misunderstanding. Will science in this field parallel the successes of the physical sciences? If such a development is to come, historical description and theoretical analysis must be supplemented by comprehensive statistical investigation.

READINGS

(See also special and general references, pages 591 and 597.)

BURGESS, ROBERT W., "The Whole Duty of the Statistical Forecaster," *Journal of the American Statistical Association*, 32 (200), December, 1937, pp. 643–648.

KOSSORIS, MAX D., "A Statistical Approach to Accident Prevention," *Journal of the American Statistical Association*, 34 (207), September, 1939, pp. 524–532.

MILLS, FREDERICK C., "Statistics and Leviathan," *Journal of the American Statistical Association*, 30 (189), March, 1935, pp. 1–11.

MITCHELL, WALTER, JR., "Industrial Wage and Hour Surveys for Management Guidance," *Journal of the American Statistical Association*, 34 (207), September, 1939, pp. 482–491.

SCHULTZ, HENRY, "Rational Economics," *American Economic Review*, 18 (4), December, 1928, pp. 643–648.

STIGLER, GEORGE J., "The Limitations of Statistical Demand Curves," *Journal of the American Statistical Association*, 34 (207), September, 1939, pp. 469–482.

WELD, L. D. H., "The Problem of Measuring Radio Coverage," *Journal of the American Statistical Association*, 33 (201), March, 1938, pp. 117–125.

CHAPTER II

THE COLLECTION OF STATISTICAL DATA

Modern statistics involves two principal aspects or divisions, the one concerned with statistical *methods*, the other dealing with the *theory* upon which these methods are formulated and critically selected and in terms of which the results of statistical analysis must be explained. Only a limited treatment of statistical theory is possible in a textbook on business statistics, however, for any comprehensive discussion of such theory leads directly into mathematics and requires an understanding of mathematical theory.

In general, the *sequence of statistical procedure* may be said to involve four principal steps or stages. First, the objective of any given analysis must be carefully defined; i.e., the purpose must be rigidly and precisely described. It is then necessary to collect significant and adequate data having implications for the defined objective. A third step involves the classification and analysis of the data thus collected to discover their pertinent characteristics. Finally, it is necessary to summarize and synthesize the results of analysis and to present the findings of the study in a form relevant to the defined objectives, to interpret the findings, and thus, so far as possible, to answer the questions framed by the definition of objectives.

Group data and individual data.—Throughout this sequence of procedure, emphasis is placed on data representing a class or classes rather than on the study of an individual item or instance. A simple definition of statistical method may be framed about this distinction between group data and individual data, which is of primary importance in understanding both the purpose and the method of statistical analysis. Statistics deals with the quantitative aspects of data; it studies groups, classes, and other aggregative forms of data to discover their significant

characteristics. Instances in which statistical methodology is essential to effective analysis might be drawn from almost every phase of modern life, from education, from business and industry, or from some other aspect of modern society.

When masses of data do not lend themselves to consideration as individual items, when their significance as a whole or as a group must be evaluated, when, for instance, it is desired to discover a measurable characteristic, the average, the range, the scatter or spread, or some other feature of the whole group, then dependence must be placed upon statistical methods.

An illustration may help to make this distinction more significant. All business firms frequently find themselves face to face with two distinct types of problems: those which require analysis on an individual basis and those which are amenable only to group analysis. In dealing with problems of personnel, for instance, a concern must frequently consider individual employees who are misfits, workers who are unhappy, unsatisfactory, and inefficient. By personal investigation, the employer may discover that such workers are physically or mentally unsound, or that financial or family circumstances are at the root of the disturbing condition. Discovery of such relationships requires individual analysis of each case, or what is frequently described as individual case study.

On the other hand, management is constantly faced with other personnel problems to which this individualistic type of analysis is obviously not applicable. For instance, questions may be raised as to the average wage paid to certain types of workers; the average length of the working day; the tendency of employees to become more or less productive with long years of continuous service; the tendency of certain types or groups of employees to have high accident, tardiness, or absence rates; or the comparative value or productivity of various types of workers in terms of education, intelligence, experience, age, length of service, marital condition, number of dependents, or other characteristics. Again, it may be important to discover what types of workers remain in the employ of the organization longest or what system of wage payment results in most efficient operation, or it may appear important to evaluate changes in

routing, production techniques, daily working hours, as well as a host of other managerial policies and practices. These and an almost endless number of similar questions cannot be answered by individual case study; their solution is dependent upon some method of analysis that permits their investigation in terms of group or aggregative data. It is for such situations that statistical methodology is prescribed.

Such situations are by no means confined to the personnel aspects of modern business. They appear everywhere throughout the whole range of business activities. For instance, the manufacturer wishes to build a foundation for forecasting future demand for his products; the retailer or wholesaler seeks to measure consumer interest in a product or group of products; the banker compares investment returns and safety for different types of securities, or analyzes customers' budgets as a basis for the extension of loans; the sales department compares different sales techniques, different territories, different advertising media; the realtor seeks measures of prevailing tendencies and trends in rents, property values, transfers, and building programs. These are but a few illustrations of the many situations which furnish fertile ground for statistical analysis.

Objectives of statistical inquiry.—The nature and objectives of such analyses may be made clearer by reference to a single specific illustrative situation. Consider the problem facing the sales manager of a small department store who seeks to discover from the reports of various departments their comparative efficiency and their needs in terms of personnel, advertising allowances, and similar requirements. He wishes to compare departments, each of which may include several employees, and possibly to contrast the sales of individual salesmen with general levels prevailing in certain departments or throughout the whole organization. To that end, he might prepare such a summary of the most important data as appears in Fig. 2·1.

What are the questions that the manager may answer by reference to such a report? Obviously, they concern such considerations as the average sale per department, the comparative cost of sales in different departments, the ratio of wages and commissions to total sales, comparisons of sales in various

departments in this week with those of other weeks, averages of the number and amount of sales per employee, comparisons of sales results with absenteeism and tardiness, and trends of business in the various departments. It will be noted that the answers refer to group or aggregative conditions and that the data are in themselves quantitative in character, being based upon measurements in terms of specific units. The comparisons made are group comparisons, and all the questions are of a type that requires some sort of statistical methodology.

Week ending Apr. 24, 1937.						
Department	Number of salesmen	Number of sales	Amount of sales	Number of absences	Number of times tardy	Wages and commissions
A	3	110	$554.31	1	0	$76.25
B	3	240	468.20	0	0	60.00
C	2	12	640.90	0	1	91.70
D	4	65	883.29	0	0	72.00

FIG. 2·1.—Summary of Departmental Reports.

This illustration, of course, is distinctly a selected and simplified one, but it suggests the essential objectives of statistical inquiry in business. Such inquiry seeks to discover significant characteristics of grouped or aggregative data. It involves collection and classification of pertinent facts, as do all sciences, but it also describes methods of analyzing and synthesizing these data to bring out certain aggregative characteristics that cannot be conveniently and reliably inferred from individual items.[1]

SOURCES OF DATA

Statistical analysis necessarily begins with the gathering and compiling of data to be analyzed. In a given concern, this task may be the special duty of one division or branch, or it may be simply one of the many functions of the general management, accomplished by means of routine reports from individual

[1] For definitions of statistical terms, see Albert H. Kurtz and Harold A. Edgerton, *Statistical Dictionary*, New York, John Wiley & Sons, 1939.

departments. Data thus secured are described as *internal reports*. In more extensive investigations, reaching beyond the limits of the individual concern, the discovery and compilation of data may necessitate sending out field workers to make direct inquiries, or it may involve collecting questionnaires from those who have the desired information. In other instances, data to be analyzed may be secured from reports, bulletins, periodicals, and similar publications, or from business services. All such data, secured from sources outside the individual firm, are regarded as *external* reports. They may be secured from either *primary* or *secondary sources*. A primary source is one in which the data are gathered and released by the same organization; a secondary source is one in which data are released by an organization other than that by which they were collected.

Secondary sources.—Within the confines of an individual business and in certain types of external statistical analysis (notably studies of consumer preferences, marketing areas, marketing devices, local shifts in purchasing power, and the like), recourse must be had to original sources. In such cases, reliance must be placed upon the services of interviewers, or questionnaires must be utilized to secure the desired information. However, a vast amount of statistical data with respect to general business conditions is available, already collected and partially analyzed by a variety of reporting agencies. Such agencies and their publications are generally referred to as *secondary sources* because those who use them do not generally have access to the original sources from which the data have been drawn.

A comprehensive list of the agencies that supply such data together with a classification of the information they make available would amount to a fair-sized book in itself. These sources are of many types; they include governmental divisions, departments, bureaus, and services; trade associations; trade publications; private business and statistical services; specialized business newspapers; and many less important types of reporting agencies. The student of business statistics will rely almost entirely upon such sources for the data used as illustrative and problem materials, and they are almost indispensable

to private business. For this reason, it is necessary to become familiar with the most important ones. A brief summary, classified according to sources, includes:

I. International organizations:

The League of Nations:
Monthly Bulletin of Statistics.
International Statistical Yearbook.
International Labour Review.
The International Labour Office:
The I.L.O. Yearbook; Vol. II: Labour Statistics.

II. National governments:

Germany:
Statistisches Jahrbuch für das deutsche Reich.
France:
Bulletin de la statistique générale (annual).
England:
Statistical Abstract (annual).
Canada:
Monthly Review of Business Statistics. Department of Trade and Commerce, Dominion Bureau of Statistics.
United States:
Department of Commerce:
Census reports. (Major divisions include: 1, population; 2, unemployment; 3, agriculture; 4, manufactures; 5, distribution.)
Survey of Current Business (monthly), most frequently used of all sources.
Statistical Abstract of the United States (annual).
Commerce Yearbook (Vol. I covers foreign and domestic trade of the nation; Vol. II presents trade of foreign nations) (annual).
Commerce Reports (weekly).
Summary of Foreign Commerce of the United States (monthly).
Market Data Handbook of the United States (annual).
Market Research Agencies (annual).
Department of Labor:
Monthly Labor Review (wages, hours, working conditions).
Labor Information Bulletin (monthly).
Mimeographed weekly releases on wholesale prices.
Bulletins on employment, wholesale prices, retail prices in specified cities, construction, costs of living.
Department of the Interior:
Mineral Resources of the United States (annual).
Special bulletins on mine accidents, metals, and minerals.

Department of Agriculture:
   Monthly Crops and Markets.
   Yearbook of Agriculture.
Treasury Department:
   Annual Report of the Comptroller of the Currency (financial, monetary, and banking statistics).
Federal Reserve Board:
   Federal Reserve Bulletin (monthly).
Federal Reserve Banks:
   Regional bulletins (monthly).
Central Statistical Board: special studies (a new agency).

III.   State governments:
   Auditor's or treasurer's reports, variously labeled (public finance).
   State universities: bulletins of bureaus of business research or of schools of commerce, finance, and business administration.
   State planning boards: special studies of population, income, marketing, living costs, and others.

IV.   Commercial and financial newspapers and periodicals:
   Commercial and Financial Chronicle (weekly).
   Bradstreet's (to 1933).
   Business Week.
   Dun's Review.
   Barron's (weekly).
   Wall Street Journal (daily).
   New York Journal of Commerce (daily).
   Chicago Journal of Commerce (daily).

V.   Trade publications (the list is typical rather than selective):[1]
   Printers' Ink (advertising lineage).
   Economic World (cotton).
   Iron Age.
   Engineering and Mining Journal.
   Railway Age.
   Northwestern Miller.
   Engineering News-Record.
   Oil, Paint, and Drug Reporter.
   Sales Management.

VI.   Trade associations (typical rather than selective):
   Committee on Public Relations of the Eastern Railroads.

[1] For a more extensive list, see J. R. Riggleman and I. N. Friskee, *Business Statistics,* New York, McGraw-Hill Book Co., 1932, pp. 322 ff.

Bureau of Railway Economics.
National Credit Men's Association.
Iron and Steel Institute.
American Petroleum Institute.

VII. Private statistical services:
National Bureau of Economic Research (prices, business-cycle studies, etc.).
National Industrial Conference Board (working conditions, living costs, etc.).
Poor's, publishing *Poor's Corporation Manual.*
Moody's, publishing *Moody's Corporation Manual.*
Standard Statistics Company.
Roger Babson, publishing *Babson's Reports.*
Brookmire Economic Service.
Harvard Economic Service.

VIII. Specialized reporting agencies, typified by:
R. L. Polk and Company (new car registrations).
Audit Bureau Corporation (publication circulations).

Reference may also be made to two rather widely used general sources which, although they do not confine themselves to business data, are frequently convenient. They are the *Statesman's Yearbook* published annually by The Macmillan Company, and the New York *World-Telegram's World Almanac,* also an annual.

Especial caution is essential in the use of data obtained from secondary sources, if misinterpretation and erroneous conclusions are to be avoided. It is important to take into account any possible bias or prejudice that may characterize the reporting agency. The purpose for which the agency was established and is maintained is probably the best guide as to possible bias. It is essential to ascertain the nature and adequacy of the sources from which the reported data are drawn as well as the methods used in their compilation. Some data represent the result of censuses which have covered the entire field, as is true of most of the population data released by the Department of Commerce. In other cases, sampling is resorted to, and it is pertinent to inquire how extensive and how representative of the whole the sample is. In general, the sample may be regarded

as satisfactory only if it contains all the relevant characteristics of the whole population in proportion to their relative importance in the whole.

Care must be exercised, also, in noting precisely what the data are assumed to represent, what limits define them, and what related conditions they may exclude. Most monthly data on employment in the United States, for instance, refer only to employment in manufacturing industries or in manufacturing and a few selected non-manufacturing industries. To assume that they represent employment in general would involve an inexcusable error. Numerous illustrations of serious misinterpretations of this sort appear regularly in newspaper reports and public addresses. For instance, a recent reference was made to the most frequently used index of production as a measure of all economic activity in the United States, including wholesaling, retailing, and construction, none of which were then included as elements in the current measure.

The measures in which data are expressed must be closely regarded. Are they expressed in absolute terms, such as pounds, tons, bushels, or bookkeeping entries, or are they relatives, such as percentages or ratios? If they are relatives, with what are they compared, what is their base, and what time unit is involved? How are the measures defined; i.e., what are ton-miles, commercial failures, industrial disputes, accessions, industrial accidents, wholesale prices, semi-manufactured articles, or other units designated by the titles of tables and summaries? If the data are to be intelligently used, all these questions must be correctly answered, and the search for correct answers is one of the penalties that must be paid for the convenience of secondary data. Reliable sources may be counted upon to exercise care in labeling their data and to provide such information as to their derivation, meaning, and limitations as will make them most useful.

If possible, of course, it is always advisable, even if not so convenient, to get to the primary source. There, terms are likely to be more carefully and thoroughly defined, and errors of transcription may be avoided. For it must be recognized that any agency, however laudable its intentions, may make

mistakes, so that every step away from the primary source increases the likelihood of error in the data.

Primary sources: use of questionnaires or schedules.—If a questionnaire is used to secure data, it is necessary to formulate the questions with great care, and, if the study is at all complicated, it may be necessary to supplement the questionnaire with detailed instructions. In certain types of inquiries, where those from whom information is sought may be reluctant or unable to answer necessary questions, interviewers may be required to secure the desired information. It is customary, under such conditions, to simplify the reporting forms and to refer to them as enumerators' "schedules" rather than as questionnaires. In such reports, major dependence is placed upon the interpretation of significant facts by interviewers instead of upon long detailed instructions designed to aid untutored respondents in preparing their answers. Thus the "schedule" is a less explicit and somewhat abbreviated form of the more elaborate and explanatory questionnaire. Such a schedule is illustrated in Fig. 2·2, which represents a form used by an office supply house to discover simple information as to the past purchases of envelopes and possibilities for future additional purchases by business and industrial concerns. References to size numbers and types would not be effective in a questionnaire submitted to inexpert users, but the salesmen who fill out this schedule have a clear understanding of these features. Hence greater detail is unnecessary, and the editing and recording of the returns are simplified.

The use of either schedules or questionnaires necessitates extensive preliminary preparation. A primary step requires careful consideration of the exact nature of the desired information: a clear-cut statement of the problem. Suppose, for instance, that a study is to be made of the incomes received by certain types of workers in a given city. It will be necessary, first, to define precisely the sort of employee to be included in the study. If the results are to be used in planning a sales campaign for a new consumers' good, for instance, coverage of a relatively small number of typical workers might be regarded as sufficient. But it would be necessary to define the type in

detail, in terms of occupation, distinctive characteristics, nature of employing plant or organization, and other such features. If the inquiry is to be limited to factory workers, for instance, it is still necessary to define a "factory" and to distinguish factories, small shops, and service establishments. This classification might be attained by reference to a classified list of business

Date_____   Salesman_____

Territory No._____   City_____

Firm name_____

Address _____

Business Classification_____

Envelopes purchased (all sources) in 1939:

Type	Size	Special features
————	————	————————————————————
————	————	————————————————————
————	————	————————————————————
————	————	————————————————————
————	————	————————————————————
————	————	————————————————————
————	————	————————————————————

Special types needed_____

Unusual requirements _____

Remarks:

Fig. 2·2.—Enumerator's Schedule.   Survey of envelope purchases and requirements.

firms from which those to be included can be selected. The inquiry would then be directed toward the determination of incomes among these selected workers.

Difficulties might arise because of the various ways in which wages are reported, and account would have to be taken, therefore, of hourly or daily wage rates, the number of hours regularly worked, and the rate of pay and usual extent of overtime. It

would certainly be desirable to extend the inquiry over a long enough period so as not to reflect merely the low wages characteristic of a slack period. For this reason, it is likely that records of incomes for several months or years would be necessary.

In other instances, it might be worth while to have information about the total number of workers and the total amount of wages paid, so as to discover the average wage. It might also be useful to include data on the ages of workers, their nationalities, marital status, and the numbers of dependents. Perhaps facts as to education, training, and experience might be pertinent in defining the type of worker.

It will be clear that, whether questionnaires or enumerators with schedules are to be utilized, the purpose of the inquiry must be carefully defined. Questions must be so worded as to be definite and readily answerable. Leading questions, i.e., those suggesting an answer, are of course to be avoided, as are those permitting misunderstanding or others that might prove offensive. Naturally, the details of such preparation vary widely from one project to another, depending on whether personal interviews or records of business firms or governmental bureaus are used. Frequently, assistance may be secured from earlier studies of a similar nature.

When the problem has been precisely stated and suitable questions have been prepared, attention must be given to the method by which the proper candidates for questioning may be reached. In some cases, a random sample may be taken; in others a stratified sample (see Chapter VII) may be required; in still others, all members of a particular group may be queried. The point is that careful planning must insure that those and only those who are the subjects of this investigation shall contribute to its findings. If, therefore, dependence is placed on a sample, it must be truly representative. Sometimes names may be taken from a telephone directory or from a registry of voters, or selected city blocks may be canvassed, or automobile licenses or drivers' license numbers may be drawn. None of these methods is universally applicable. They are but examples of possible devices. Any method that secures an adequate and

STATE PLANNING BOARD

A. Trading Center:
1. In what town do you do most of your trading? _____ _____
2. How many miles do you live from that town? _____
3. Why do you prefer this town?
   (Check) _____ a. Nearness   _____ d. Better roads
       _____ b. Credit   _____ e. Lower prices
       _____ c. Better stock of goods _____ f. _____
4. Did you live on this farm in 1920? _____

B. Purchases
5. Name of town where following goods are usually purchased by personal visit to stores:

GOODS	TOWN		If any of these goods are usually purchased by mail order, check below	
	1934	1920 (If living on this farm)	1934	1920
a. Groceries				
b. Drugs, Medicines				
c. Lumber, Cement				
d. Woman's coat or dress				
e. Woman's shoes				
f. Man's suit				
g. Man's overalls				
h. Farm Machinery				

6. Do you usually purchase any of the above goods by mail order? If so, check items above.

C. Sales
7. Name of town where following products are marketed:

PRODUCTS	TOWN	
	1934	1920 (If living on this farm)
a. Hogs		
b. Cattle		
c. Grain		
d. Eggs—Poultry		
e. Cream		

D. Banking
8. Do you carry a checking account in a bank? _____
9. If so, in what town? _____
10. Did you carry a checking account before the depression (1929)? _____
11. In what town? _____
12. If you do not carry a bank account, why not? (Check one)
   _____ a) Object to service charges and tax on checks
   _____ b) No local bank
   _____ c) Losses in closed banks
   _____ d) Don't need it
   _____ e) _____
13. If you have changed place (town) of banking since 1929, check the most important reason for doing so.
   _____ a) Bank closed   _____ e) Better trading town
   _____ b) Bank absorbed (consolidation) _____ f) Personal relations with bankers
   _____ c) Too hard to borrow   _____ g) Change of residence
   _____ d) Service charges   _____ h) _____

Name of enumerator _____ Date _____

Form 10B

FIG. 2·3.—Questionnaire of Committee on Business and Industry, Iowa State Planning Board.

truly representative sample may be used; suitable techniques vary with the nature of the investigation.

The accompanying inquiry used by a state planning board committee on business and industry (see Fig. 2·3) represents a fairly simple questionnaire. It is designed to secure information about the trading habits of a rural population, and the questions are pertinent to that purpose, clearly stated, and readily answerable. In order to assist the layman in understanding the type of data sought, possible answers are frequently suggested. The same device is illustrated also in the questionnaire of which a portion is shown in Fig. 2·4, which was widely circulated among those whose names appeared in the records of stock registry agencies. Under item II, for instance, several possible answers are suggested, but space is provided for different answers as well.

Market analysis.—One of the most frequent uses of questionnaires and schedules in the business field is involved in what is generally described as market analysis. Such analysis aims to discover and define distinctive and significant characteristics of a given area, thus providing a basis for planning advertising and sales programs.

Some data are available for this purpose from secondary sources. Census reports, for instance, provide information with respect to sex and occupational distribution, population origins (native or foreign born), marital status, age distribution, and related characteristics of the population. Furthermore, governmental agencies have prepared, for many sections of the country, detailed analyses of retail trade, classified according to the principal types of commodities, thus providing many clues as to the buying habits of these areas.[1]

Local studies of trading areas require the use of original sources. Enumerators are usually sent out to canvass the cities and the rural neighborhoods, the data being gathered upon

---

[1] The Market Data Section of the Marketing Division, U. S. Bureau of Foreign and Domestic Commerce, has extended these studies. See E. L. Lloyd, "Development of Retail Sales Indexes," *Survey of Current Business*, 16 (2), February, 1936, pp. 16–20, and Reba L. Osborne, "Regional Sales of General Merchandise in Small Towns and Rural Areas," *ibid.*, 16 (9), September, 1936.

November 9, 1938.

SURVEY OF STOCKHOLDER OPINION ON CAPITAL FORMATION

by

THE NATIONAL ASSOCIATION OF MANUFACTURERS

I. Do you have money available which you could invest and would like to invest in *new* securities of either new or existing productive enterprises (as distinct from government and other high grade bonds or the well-seasoned stocks of existing companies) but which you do not care to invest in such securities at the present time?

☐ *YES*
☐ *NO*

If your answer is "yes" then please answer the following questions.

II. Is your lack of willingness to make additional investment at the present time due to (*check one or more of the following*)—

☐ a. Inadequate profits being made at the present time.

☐ b. Doubt that adequate profits will be earned in the future (even if being earned now) because of—

  ☐ 1. Probable labor troubles
  ☐ 2. Probable international troubles
  ☐ 3. Existing legislation restricting industry
  ☐ 4. Possible new legislation restricting industry
  ☐ 5. Existing taxes on industry
  ☐ 6. Possible new taxes on industry
  ☐ 7. Other reasons (please specify)................................
  ....
  ....

☐ c. Even if adequate profits are earned now or in the future the government takes so much in taxes that investment is not worthwhile

  ☐ 1. Too much taken directly from the company
  ☐ 2. Too much taken directly from the individual taxpayer

☐ d. Even if adequate profits are earned the directors distribute too small a portion to stockholders.

☐ e. Government legislation places too stringent restrictions on the

  ☐ 1. Purchase of securities by individuals
  ☐ 2. Sale of securities by individuals
  ☐ 3. Issuance of new securities by corporations

FIG. 2·4.—Portion of a Questionnaire Used to Discover Reasons for Non-Investment. From "A Key to More Jobs," National Association of Manufacturers, 1939, p. 10.

suitable schedules from a predetermined sample (e.g., one out of five or ten homes may be sufficiently large). Maps showing the trading areas for different types of merchandise may be prepared from such an investigation. Figure 4·21 (page 66) is an illustration. Cf course, the purpose of analysis may vary to include discovery of changes in buying habits, choices among brands or substitute goods, preferences in advertising media, and numerous other significant characteristics of the market.

Effective analysis and intelligent interpretation of these data require the use of many of the statistical techniques to be described in subsequent chapters. Discovery and measurement of trends, seasonal fluctuations, cyclical influences, covariation or correlation among related market characteristics, and the use of relatives to represent measures of variation are some of the numerous statistical procedures which are constantly utilized. In general, it is the basic purpose of business statistics to provide an adequate description of productive and distributive· processes, including investment and other financial operations, so that the most effective adaptation of business resources to social needs may be made.

READINGS

(See also special and general references, pages 591 and 597.)

BAIN, READ, "Stability in Questionnaire Response," *American Journal of Sociology*, 37 (3), November, 1931, pp. 445–453.

ELMER, M. C., *Social Research*, New York, Prentice-Hall, 1939, Chapters XVI, XVII, XX.

LUNDBERG, G. A., *Social Research*, New York, Longmans, Green & Co., Chapters VI, VII.

RUCHMICK, CHRISTIAN A., "The Uses and Abuses of the Questionnaire Procedure," *Journal of Applied Psychology*, 14 (1), February, 1930, pp. 32–41.

SHUTTLEWORTH, FRANK K., "A Study of Questionnaire Technique," *Journal of Educational Psychology*, 22 (9), December, 1931, pp. 652–658.

YOUNG, P. V., *Scientific Social Surveys and Research*, New York, Prentice-Hall, 1939, Chapters III, IV, VIII.

EXERCISES

1. Under what conditions is statistical analysis typically desirable or necessary?

2. Distinguish internal and external reports, and cite illustrations of each type.

3. Distinguish primary and secondary sources of statistical data, and cite an illustration of each.

4. Assume that you have been given the assignment of finding what radio programs are of most interest to the citizens of a small town of about 2,500 population. Prepare a questionnaire to be used for that purpose. Then assume that you will use interviewers to gather the data and prepare an interviewer's schedule for this purpose.

5. Assume that, for the purposes of Exercise 4, you wish to restrict analysis to a sample. Outline the method you would pursue in order to secure a truly representative sample.

CHAPTER III

CLASSIFICATION AND TABULATION

Editing data.—After data have been collected, it is necessary to edit and classify them, and it is frequently desirable to summarize them in convenient tabular form for presentation or for further analysis. Editing discloses deficiencies in reporting, and it may indicate some of the immediate limitations of the data. Information gathered by enumerators is usually subject to less error than that which is reported directly on questionnaires, but even that collected by means of the most carefully prepared schedules generally requires careful editing.

The editing process varies somewhat with the nature of the data and the projects for which they have been collected, but a few principles of almost universal application deserve mention. Returns should be checked carefully to insure that they are complete, and those that do not meet this test should be separated from the remainder. Replies should be checked for consistency, i.e., to see that they do not contain contradictions within themselves which indicate some error in reporting, and all inconsistent returns should be segregated. The test of reasonableness should be applied, for returns cannot be accepted if they describe situations known to be highly improbable or impossible. An attempt should be made to correct or adjust the defective reports. If that is impossible, they must be returned to the original sources of information, discarded, or replaced with other reports.

In many cases, certain minor computations remain to be made before data are recorded, and it is frequently convenient to "code" the data before subjecting them to further manipulation. Many types of codes may be used. In one of the simplest, answers are designated as numbers or letters, so that they may be briefly recorded within limited space. Thus a given

TABULATION SHEET

DATE OF TABULATION March 1, 1938

TABULATED BY D. L.

FIG. 3·1.—A Tabulation Sheet. Data: A generalized personnel audit of 100 firms in Minneapolis, 1937.

answer to a question or combination of questions may be numbered *7*, while a contrary response may be coded as *8*.

Classification of data.—When the data have been edited, they are ready to be classified for further analysis. The most common bases for such classification are. (1) chronological, (2) quantitative, (3) geographic, and (4) qualitative.[1] Sometimes this classification is accomplished most easily by copying the data directly upon a sheet such as that illustrated in Fig. 3·1. In others, data are transcribed upon small cards which may then be sorted and shifted about in various classifications. Where large forms are used in reporting extensive data, more complicated tally-sheets may be set up and items checked in their appropriate columns and rows.

The variety of tally-sheets and work-sheets that is necessary to summarize and analyze the data included within the wide range of studies in business statistics makes it quite impossible to describe these devices in detail. The form such sheets will take in any given study depends essentially upon the nature of the data and the purposes of the investigation. Figure 3·2 illustrates one of several convenient forms for assembling monthly data over a period of years.

Machine tabulation.—In dealing with large numbers of items, complicated mechanical aids, generally known as sorting and tabulating machines, are used. Such devices transfer the original data from questionnaires, schedules, reports, or other sources to "code cards" such as those illustrated in Fig. 3·3, each of which is punched to indicate the particular characteristics of the item it represents. Cards are of uniform size, and they generally have either 45 or 80 columns of digits. In their simplest form, these cards contain no captions and may be readily adapted to a variety of uses by designating the columns to correspond with the number of answers on a particular questionnaire. Where cards are prepared for particular and extensive types of analysis, they may have printed designations for each column, as is illustrated by the second of the two cards in Fig. 3·3. It is apparent that any characteristic or combination of features

[1] For a detailed discussion of editing and tabulation, see Bruce D. Mudgett, *Statistical Tables and Graphs*, Boston, Houghton Mifflin and Co., 1930, pp. 8 ff.

FIG. 3·2.—Monthly Data Sheet, by permission of Codex Book Co., Inc., Norwood, Mass.

that can be represented by numbers may be recorded by appropriate punching of such cards, and their only limit is the number of columns and digits.

The first step in machine calculation involves the transfer of data from the original records to the tabulating cards by punching. The code cards are then prepared for sorting and

Fig. 3·3.—Sample Tabulating Cards.

counting. In the machine counting process the totals and other desired values are readily available.

The most common of the tabulating machines in use are the Hollerith machines (a product of International Business Machines Corporation, and the type using the cards shown in Fig. 3·3) and the Powers (produced by Remington-Rand Business Service, Inc.). In each case, the essential equipment con-

sists of one or more key punches, a sorting device, and a tabulating machine. In one type, the tabulating device records the desired values on a separate sheet; in the other, the results of the tabulation are read directly from the machine. The Hollerith machine makes use of electrical circuits in its tabulation and may be readily manipulated to effect various types of calculations by changing the electrical connections in its control panel. The Powers machine utilizes a mechanical tabulating device for this purpose. In each case, the equipment is operated by electric motors, so that a vast amount of human effort is avoided, the work is greatly speeded, and errors are reduced. From 2,000 to 3,500 cards may be punched in one day by a single operator, and sorting is accomplished at the rate of 250 to 400 cards per minute. Tabulation requires about twice as much time as sorting.

Tabular presentation.—When the data have been classified, it is customary to prepare them for presentation in the form of a table or tables, so that they may be clearly understood and properly utilized. The basic problem is to secure the maximum of clarity and to avoid misinterpretation; several simple rules may be noted in this connection.

It is important that the title of the table clearly defines the nature of the data, setting out their limits in unmistakable terms. If there can be any questions as to the nature of the units in which the data are expressed, then these questions should be answered fully in the title or in subtitles. The source of the data, unless it is original, should be so described as to permit verification by an interested reader; the usual method is to indicate the magazine, bulletin, journal, or other printed source by name, volume, number, date, and page.

One of the major purposes of tabulation is to show relationships among the data, indicating which are subordinate and which are of coordinate importance. This result is obtained in a variety of ways, one of the most common being the use of distinctive sizes and styles of type; another involves designation of the more important classifications in column headings, and the less important cross-classifications in the left-hand column, which is generally referred to as the "stub."

Detailed characteristics of such tables depend upon the nature of the inquiry, but Table 3·1 is an illustration sum-

TABLE 3·1

ANNUAL RENTS OF TWO-STORY, ONE-FAMILY DWELLINGS IN —— CITY, 1936–1939

Source: Data selected from an unpublished planning-board study

(Read upper class limit as "up to but not including" the stated figure.)

Annual rent paid	Number of dwellings in each year			
	1936	1937	1938	1939
$400–$500 —	31	33	31	30
500– 600 —	45	51	56	55
600– 700 —	60	65	62	66
700– 800 —	40	31	35	33
800– 900 —	11	10	9	10
900–1000 —	2	2	1	3
Total.........	189	192	194	197

marizing the facts discovered by means of a questionnaire like those previously described (pages 19–25). The investigator in this particular study selected from the returns those meeting the following qualifications: (1) two-story single-family dwellings in a typical residential district; (2) dwellings held fairly constant in value by minor improvements, such as redecoration and repairs; and (3) dwellings constantly occupied except for short intervals when tenants changed dwellings for which an annual lease-contract was required.

In this case, it happens that the data are readily classified according to the amount of rent, and this classification forms the subject-matter of the stub of the table. The table caption represents a cross-classification, also simple, and based upon the year involved. Another tabular arrangement of somewhat different data is shown in Table 3·2, which presents an urban-rural distribution of the population of the United States, as of

1930. It will be noted that the "heading" or caption is subdivided by "boxing" the titles of the columns.

TABLE 3·2

DISTRIBUTION OF POPULATION IN URBAN AND RURAL TERRITORY, 1930
Source: United States Census of Population, 1930.

	Number of places	Population	
		Number	Per cent of total population
United States.....................	......	122,775,046	
Urban territory................	3,165	68,954,823	56.2
Places of 1,000,000 or more....	5	15,064,555	12.3
Places of 500,000 to 1,000,000..	8	5,763,987	4.7
Places of 250,000 to 500,000...	24	7,956,228	6.5
Places of 100,000 to 250,000...	56	7,540,966	6.1
Places of 50,000 to 100,000....	98	6,491,448	5.3
Places of 25,000 to 50,000.....	185	6,425,693	5.2
Places of 10,000 to 25,000.....	606	9,097,200	7.4
Places of 5,000 to 10,000......	851	5,897,156	4.8
Places of 2,500 to 5,000.......	1,332	4,717,500	3.8
Rural territory................	13,433	53,820,223	43.8
Incorp. places, 1,000 to 2,500...	3,086	4,819,430	3.9
Incorp. places under 1,000.....	10,347	4,363,605	3.6
Other rural territory..........	......	44,637,188	36.4

Tabular presentation of this sort may involve twofold, threefold, or fourfold subdivision and cross classification, or it may be even more complicated. The arrangement may take many different forms. The only rule that can be stated is that it must be so designed as to present the data in an unambiguous and convenient form. Many excellent examples of tabulation may be found in the publications of the Federal Census Bureau as well as in the *Statistical Abstract of the United States* and such current publications as the *Survey of Current Business* and the *Monthly Labor Review*.

The array.—A simple form of arranging data for purposes of comparison or presentation and for certain types of statistical analysis is that known as the *array*, in which the items are listed in the order of their size, usually beginning with the smallest and ending with the largest, although the reverse order is sometimes preferable. For example, suppose that the monthly electric bills of a small store are found to run as follows (all figures in dollars): 16.50, 12.75, 9.00, 12.50, 14.00, 15.00, 18.00, 13.50, 14.25, 11.50. It is clearly easier to estimate the relative size and variability of these items if they are listed as an *array*, as follows: 9.00, 11.50, 12.50, 12.75, 13.50, 14.00, 14.25, 15.00, 16.50, 18.00. Inspection of these figures shows clearly the total *range*, i.e., the spread from lowest to highest or from smallest to greatest, and some idea of their average is suggested by the items near the center of the array.

The frequency distribution.—In most statistical analysis, data are classified on a quantitative basis, that is, according to size or magnitude. When data have been so classified, and the numbers of items in each class noted, they are said to represent a *frequency distribution*. Table 3·2 presents such a distribution. The cities are there classified according to size, and the number of cities in each class has been counted and entered in column 2. The first class includes cities of 1,000,000 or more, the second class cities having populations from 500,000 to 1,000,000, and the last class cities of 2,500 to 5,000 population. It may well be noted that this table, one of a type often encountered, is featured by unequal *class intervals* and represents an *open-end* distribution. Thus, the *class intervals* vary from 1,500 (the range from 1,000 to 2,500) to 500,000 (the range from 500,000 to 1,000,000). Although such classes are frequently convenient as means of presenting data, they are not so useful if further calculations are to be made, if, for instance, an average for the whole is to be found. Further, this difficulty is increased when an *open-end class*, such as that designated "Places of 1,000,000 or more," is included. Hence, if data are to be subjected to further analysis, it is well to avoid these features in classification.

A more simple frequency distribution is illustrated in

Table 3·3, where 40 employees of a business concern were classified according to their ages in years and fractions of years. The raw data of individual ages were subjected to classification, so that the ages of the 40 individuals appear as 9 age groups. The most common characteristics of the frequency distribution may be readily observed from this simple illustration. It will be noted, for instance, that the classes are all inclusive; there are no items representing individual ages outside their limits. At the same time, the classes are mutually exclusive; they do not overlap, and there is, therefore, no question as to where each magnitude should be tabulated.

This procedure is made possible by the careful designation of what are known as *class limits* (symbol L), which define or delimit each of the classes and indicate the range of items to be tabulated in each class. Thus the lower limit (L_1) of the first class is 15 years, and the upper limit (L_2) of that class is 20 −, or just under 20 years. For the second class, these limits are 20 and 25 −. It is customary, in describing such limits, to say that each class extends from its lower limit up to, but not including, the lower limit of the next class. Thus the first class extends from 15 years up to, but not including, 20 years; the second, from 20 years up to, but not including, 25 years; and so on throughout the tabulation. Sometimes, this idea is expressed by describing the upper limit of each class as including a decimal fraction that places it just below the lower limit of the next class. Thus the upper limit of the first class might be described as 19.9 or 19.99 years (depending upon the number of decimal places to which individual items were recorded), and that of the second class might be 24.9 or 24.99 years, thus recognizing the fact that the magnitude *19* includes individuals who are all the way from 19 up to, but not including, 20 years of age. The use of the minus sign after the upper limit, when the latter is made identical with the lower limit of the next class, illustrated in the example, accomplishes the same result.[1]

---

[1] In some cases, items may fall exactly upon the class limits, and, if they are entered in either of the adjacent classes, an error will be thereby introduced. Such errors may be offset by entering any such item as a half frequency in each of the adjacent classes.

The *mid-point* or *class measure* of each class (symbol m) is strictly the average of the exact limits, that is,

$$m = \frac{L_2 + L_1}{2}$$

In the example illustrated by Table $3 \cdot 3$, the measure of the first class is the average of 15 and 20 $-$, or 17.5. The same measure (approximately) is obtained if the limits are regarded as 15 and 19.99. Similarly, the m of the second class is $(25 + 20-) \div 2$, or 22.5, and the other measures have been calculated in the same manner and are noted in column 2 of the table.

TABLE 3·3

THE FREQUENCY DISTRIBUTION

Data: Ages of 40 employees of a retail store.

1 Class limits		2 Mid-point or measure	3 Fre- quency	4 Cumulative frequencies		5 Frequencies as per cent	6 Cumulative, $f, \%$	
L_1	L_2	m	f	Σ_1	Σ_2	$f, \%$	Σ_1	Σ_2
15	20 $-$	17.5	2	0	2	5	0	5
20	25 $-$	22.5	6	2	8	15	5	20
25	30 $-$	27.5	10	8	18	25	20	45
30	35 $-$	32.5	8	18	26	20	45	65
35	40 $-$	37.5	5	26	31	12.5	65	77.5
40	45 $-$	42.5	4	31	35	10	77.5	87.5
45	50 $-$	47.5	2	35	37	5	87.5	92.5
50	55 $-$	52.5	2	37	39	5	92.5	97.5
55	60 $-$	57.5	1	39	40	2.5	97.5	100

Continuous and discrete series.—When ages are recorded in years and fractions of years, as in the illustration, the series thus obtained may be regarded as being practically continuous, for ages range through almost infinitely small variations. In many other types of data, however, items appear as discrete integers, and the series or distribution appears as discontinuous. In gen-

eral, continuous series result from a measuring process, whereas discontinuous series follow from a counting process. Thus, for instance, if railroad passenger cars are classified according to the number of passengers counted at a given time, there would be no fractional items. In such cases, class limits are generally stated in terms of the integers that define each class. A first class might be 15 to 19 passengers, a second 20 to 24 passengers, with others similarly defined. There would be no serious problem of classification, since no item could fall between adjacent limits. But for interpolation, $L_1 = 14.5$, $L_2 = 19.5$, etc.

In discontinuous series, it is worth noting that the theoretical limits of the classes just described are 14.5 to 19.5, 19.5 to 24.5, and so on throughout, for each integral item actually represents an area extending from $\frac{1}{2}$ unit below to $\frac{1}{2}$ unit above the integer. Thus the area (in the whole range) described as 15 runs from 14.5 to 15.5, and that described as 20 includes the distance from 19.5 to 20.5. This conception of the nature of the integers and of the classes they define will be found helpful in several types of statistical analysis to be presented later. Here it need only be described, and it may be mentioned that, in most types of tabulation and classification, measures are originally made to a certain convenient measure of exactness. Since the data are, as a result, only approximations in convenient denominations, theoretical limits are logically set at the mid-point between such denominations. The use of the convenient integral limits, i.e., 15 and 19 as contrasted with 14.5 and 19.5, does not introduce an error, for the mid-point is not changed: the average of 15 and 19 is the same as that of 14.5 and 19.5.

Few hard and fast rules can be given for the selection of class limits, principally because much depends on the purpose of the investigation and the type of analysis to be undertaken. In general, however, it is desirable that limits be so selected that their averages, the mid-points or class measures, are convenient for manipulation. More important, the limit must be such that the mid-point fairly represents the class as a whole. Hence they must not be so selected that data "bunch" close to one or the other limit, since this condition prevents the mid-point or class average from being truly representative. It is desirable,

also, that they be such that no item falls directly on a class limit, for this condition requires special adjustment.

In general, it may be said that, where the data are regarded as continuous, most convenient practice describes the upper limits of each class as identical with the lower limit of the next class and it describes the class as extending from the lower limit up to but not including the upper limit, i.e., 15 to 20 − ; 20 to 25 − ; etc. In discontinuous series of discrete items, limits are most commonly defined as the smallest (L_1) and the largest (L_2) items respectively that are to be included in the class, i.e., 15 to 19; 20 to 24; etc., but interpolations require 19.5, etc.

The nature of the data and the purpose of the investigation necessarily determine the number of classes to be distinguished. Use of a few classes, four or five, for instance, presents a simple, broad survey of the distribution, but such an arrangement sacrifices accuracy in calculations. The use of many classes, on the other hand, tends to increase accuracy but prevents a comprehensive view of the nature of the distribution.[1]

Cumulative frequencies.—Column 3 of Table 3·3 requires little explanation, for it merely lists the numbers of items or frequencies in each class. Thus, it appears that there were 2 persons in the first age group, 6 in the second, and 10 in the third. The sum of the items in this column is obviously the total number of items in the distribution; it is generally designated as N.

The next column, described as " Cumulative frequencies," requires more attention. Its two subdivisions are labeled Σ_1 and Σ_2, respectively, and they cumulate the total frequencies

---

[1] For a discussion of the theoretically desirable size of classes for frequency distribution, see Herbert A. Sturges, "The Choice of a Class Interval," *Journal of the American Statistical Association*, March, 1926, pp. 65–66. On the basis of binomial distributions, as discussed later, the number of items (N) is related to the appropriate number of classes (C) as follows:

N	C	N	C	N	C
16	5	128	8	1024	11
32	6	256	9	2048	12
64	7	512	10	4096	13

Or, in general,

$$N = 2^{(c-1)}; \text{ and } C = 1 + 3.322 \log N$$

from the beginning of the distribution to L_1 and L_2 of each class, respectively. Thus Σ_1 for each class is the total number of items up to the beginning or L_1 of this class. For the first class, therefore, Σ_1 is obviously 0. For the second class, it is the number of items in the first class, or 2. For the third class, it is the sum of the items in the preceding two classes, or $2 + 6 = 8$. The other cumulative total, Σ_2, is the sum of frequencies up to the upper limit, L_2, of each class. Thus, for the first class, it is the number of items in that class, or 2. For the second class, it is the total of the items in classes one and two, or 8. For the third class, it is the total of the items in the first three classes, or 18. Stated in another way, these cumulatives represent the number of items whose measure is less than the limit to which the cumulatives apply. Thus the Σ_1 of the third class is the number of persons whose ages are less than L_1 of that class, or 25 years. Similarly, the Σ_2 of that class represents the total number of persons whose ages are less than L_2 of that class, or 30 years.

Because of this characteristic of these cumulatives, they are frequently described as the " less than " cumulatives. If data are cumulated in reverse order, beginning with the final class and extending toward the beginning class, as is frequently done, the resulting cumulatives are called the " more than " cumulatives, since they express the numbers of items that are greater in measure than the class limits with which the individual cumulatives are associated. The " less than " cumulatives are the more common and generally the more convenient to use.

It will be noted from this description and from Table 3·3 that the cumulatives associated with lower class limits, those labeled Σ_1, necessarily begin with zero, and successive items in this column are identical with the Σ_2 of the preceding class. Because of this repetition, the Σ_1 column is often omitted, and a single column, actually Σ_2 but designated merely as Σ, is used to represent the cumulatives. However, it may frequently be found convenient to have both cumulative columns expressed for purposes of interpolation to be described in later discussions.

Column 5 of the table is, as its label indicates, the expression of the frequencies in each class as percentages of the total num-

ber of items in the distribution. Thus the first class, which includes 2 persons, appears as $\frac{2}{40}$ or 5 per cent. Similarly, the second class is $\frac{6}{40}$ or 15 per cent; and the final class is $\frac{1}{40}$ or 2.5 per cent. In the next column, the cumulative percentages corresponding to the cumulative totals of column 4 are noted. In each class, Σ_1 is the Σ_1 of column 4 expressed as a percentage of the total number of items. Thus, Σ_1 for the second class in column 6 is $\frac{2}{40}$ or 5 per cent; Σ_2 for this class is $\frac{8}{40}$ or 20 per cent. As in the cumulative frequencies, common practice makes use of the Σ_2 column much more than it does of the Σ_1 column, although both may be found convenient.

Normal distributions.—Many of the distributions encountered in statistical analysis tend to approximate what is generally described as the normal distribution.[1] It is so called because it is widely believed to describe the *normal* or natural distribution of events in nature and because it closely approximates many actual distributions in a wide range of fields, including biology, psychology, physiology, and others. Its general appearance may be seen in Fig. 7·1, page 154. It will be noted that the curve is bell-shaped, indicating that the larger frequencies are clustered about its center, and the distribution is symmetrical about its central ordinate. It may be shown that the distribution represents an expression of the laws of chance or accident or probability, for which reason it is the basis of much statistical theory.

Most of the actual distributions encountered in business statistics are not symmetrical, i.e., the largest frequencies precede or follow the center or mid-point in the scale of magnitudes. Such distributions are said to be *skewed*. Sometimes skewed distributions can be made to appear approximately normal by plotting the frequencies on the logarithms of class limits or mid-points instead of on these measures themselves. When this is true, the distributions are said to be *logarithmic normal distributions*.

---

[1] The normal curve of distribution or the curve of a normal distribution, like the distribution itself, goes by many names. Normal distributions are sometimes called Gaussian distributions, normal frequency distributions, normal probability distributions, and by similar designations, and the curve is referred to as a normal frequency curve, a normal probability curve, a curve of error, and similar terms.

Summary.—It is difficult if not impossible to give precise directions for setting up all frequency distributions. Variations in the data to be analyzed call for numerous modifications in procedure. If the data are limited in number, their arrangement in the frequency distribution may result in serious inaccuracies and misrepresentations. If the data are numerous, few or large numbers of classes may be used, according to the purposes of the analysis. The very nature of this type of arrangement, which classifies the items and applies to each the measure of the mid-point of its class, loses something in accuracy, but it gains in clarity of presentation and in convenience for further analysis. The question as to the desirability of its use and the detail to which classification is carried can be answered, therefore, only in terms of the desired balance between convenience and precision. Usually, a classification involving fewer than 5 intervals is regarded as rather broad and inaccurate, while the maximum that can be conveniently manipulated is about 30 classes.

The most common difficulty encountered in the preparation and use of the frequency distribution is the tendency for data to " bunch " at certain points. Thus wages are found to center about the dollar and half-dollar marks, so that a daily wage of $2.50 or $4.00 is more common than one of $2.68 or $4.73. Under such circumstances, care must be taken to arrange the class limits so that the " bunching " falls about the mid-points of the classes, rather than at one or another extreme. If data are allowed to bunch at the extremes, the mid-points, obviously, will not be representative. If data are extremely irregular in this respect, they should be regarded as unsuitable for analysis by means of the frequency distribution.

Other methods of classification and tabular presentation will be noted in connection with various types of statistical analysis considered in subsequent chapters. Tables, however, represent only one of several methods of presenting statistical data. Attention may next be directed to another, the use of charts and graphs.

READINGS

See "Classified readings from readily available texts," pages 591–597, also:

BAEHNE, G. W., *Practical Applications of the Punch Card Method in Colleges and Universities*, New York, Columbia University Press, 1935.

CARVER, HARRY C., "The Concept and Utility of Frequency Distributions," *Proceedings of the American Statistical Association*, 26 (173A), March, 1931; pp. 33–36.

MUDGETT, BRUCE D., *Statistical Tables and Graphs*, Boston, Houghton Mifflin Co., 1930; Part I, pp. 3–60.

WALKER, HELEN M., and DUROST, W. N., *Statistical Tables; Their Structure and Use*, New York, Bureau of Publications, Teachers College, Columbia University, 1936.

EXERCISES AND PROBLEMS

A. EXERCISES

1. What is meant by "editing" the data secured from questionnaires, schedules, or reports, and why is such editing necessary?

2. What are the most commonly used bases for classification of data?

3. How are data "coded" for classification and tabulation? Why is coding desirable?

4. How do class limits for continuous series vary from those used in classifying discontinuous series?

5. The following series is assumed to represent the daily earnings in dollars of 100 workers. Tabulate them in classes having the limits $2.50–$3.50; $3.50–$4.50, etc.

10.25	7.03	8.00	4.25	9.75	4.76	7.78	7.89	4.32	4.88
5.36	5.00	4.80	5.93	3.61	4.11	4.56	6.30	5.12	6.41
4.60	3.54	5.85	6.26	4.68	5.74	3.82	3.89	7.14	7.47
6.64	4.46	6.37	9.12	3.75	6.19	5.40	4.64	5.81	5.48
4.92	6.04	4.52	9.38	5.44	7.25	5.96	5.04	5.89	7.42
5.08	7.31	3.00	6.92	5.78	8.11	6.15	5.59	6.75	5.70
5.67	8.44	5.52	4.84	6.00	8.33	6.58	6.07	6.86	5.20
7.19	5.24	6.53	3.68	7.36	6.44	8.22	6.33	5.63	4.72
5.16	3.96	7.56	6.11	5.56	7.67	8.88	6.69	5.32	4.18
4.04	6.81	8.62	7.08	4.96	5.28	6.97	6.22	4.39	6.48

6. Tabulate the following items in classes of 5–7; 7–9; etc.

12.32	9.21	8.75	17.71	14.94	9.62	9.38	16.90	11.20	9.46
7.42	7.75	10.62	10.38	10.12	7.58	17.43	12.16	16.30	15.70
15.50	16.50	7.92	8.08	7.08	15.90	12.64	8.58	10.71	14.00

13.18	5.50	13.53	6.50	18.29	14.71	19.33	11.84	12.88	10.79
16.10	10.04	11.04	7.25	12.72	13.88	12.40	14.24	10.54	12.00
12.56	14.12	9.12	13.76	9.79	11.76	11.36	15.30	12.48	8.42
9.04	10.46	9.88	13.65	12.96	10.29	9.71	18.57	14.59	12.80
17.14	11.12	13.06	18.00	18.86	10.96	8.25	14.35	10.88	20.67
20.00	11.60	11.28	9.29	14.47	16.70	8.92	11.52	11.68	12.24
10.21	14.82	9.96	11.44	13.29	12.08	11.92	9.54	13.41	15.10

7. The following data are assumed to represent two indexes that are to be compared. Each item is first tabulated in the usual manner, according to the class intervals indicated, and the two distributions are then combined to form a single double-frequency distribution.

What characteristic of the two distributions is brought out by the double-frequency distribution?

I. Original data.

Worker	Test score, Index A	Efficiency, Index B	Worker	Test score, Index A	Efficiency, Index B
A	13	19	I	22	24
B	17	16	J .	14	16
C	13	15	K	9	14
D	19	25	L	9	11
E	14	8	M	14	14
F	10	6	N	6	4
G	11	12	O	19	21
H	19	20	P	15	15

II. Each index tabulated (fill in f).

	Index A			Index B	
Class	m	f	Class	m	f
3.5– 7.5	5.5		2.5– 7.5	5	
7.5–11.5	9.5		7.5–12.5	10	
11.5–15.5	13.5		12.5–17.5	15	
15.5–19.5	17.5		17.5–22.5	20	
19.5–23.5	21.5		22.5–27.5	25	

III. Double-frequency distribution.

Index B		Index A: Class and m					
Class	m	3.5–7.5	7.5–11.5	11.5–15.5	15.5–19.5	19.5–23.5	
		5.5	9.5	13.5	17.5	21.5	f
22.5–27.5	25				/	/	2
17.5–22.5	20			/	/ /		3
12.5–17.5	15		/	/ / / /	/		6
7.5–12.5	10		/ /	/			3
2.5– 7.5	5	/	/				2
	f	1	4	6	4	1	16

8. In the following distribution, fill in the f (%) and Σf (%) columns.

Class limits, L_1 and L_2	Class mark, m	Frequency, f	Percentage frequency, f (%)	Cumulative frequencies Σf	Cumulative frequencies Σf (%)
\$2.50–\$ 3.50	\$ 3.00	5		5	
3.50– 4.50	4.00	70		75	
4.50– 5.50	5.00	125		200	
5.50– 6.50	6.00	135		335	
6.50– 7.50	7.00	90		425	
7.50– 8.50	8.00	45		470	
8.50– 9.50	9.00	20		490	
9.50– 10.50	10.00	10		500	
		$N = 500$			

ANSWERS

5. m	f	Σf	**6.** m	f	Σf
3	1	1	6	2	2
4	14	15	8	12	14
5	25	40	10	24	38
6	27	67	12	25	63
7	18	85	14	17	80
8	9	94	16	10	90
9	4	98	18	7	97
10	2	100	20	3	100

7. Index A: $f = 1, 4, 6, 4, 1$. Index B: $f = 2, 3, 6, 3, 2$.
8. $f (\%) = 1, 14, 25, 27, 18, 9, 4, 2$.

B. Problems

9. In the table below, the years of service and efficiency ratings of 20 employees are given.

(*a*) Prepare a frequency distribution of each set of data.

(*b*) Prepare a double-frequency distribution employing the same class limits as above, and taking years as X and ratings as Y.

Employee	Service (in years)	Rating	Employee	Service (in years)	Rating
A	1	5	K	6	9
B	9	6	L	7	4
C	8	8	M	1	2
D	3	8	N	1	3
E	3	6	O	3	8
F	2	7	P	1	6
G	4	5	Q	2	5
H	5	6	R	2	3
I	5	4	S	4	4
J	6	5	T	2	7

10. Summarized below are the individual sales, in dollars, for 100 salesmen. Organize these data in the form of a frequency distribution based on the magnitude of sales, and prepare a table showing the classes, the class measures, frequencies in each class, and cumulative frequencies. Add a column to your table thus prepared in which cumulative frequencies are calculated from the large classes to the small classes (m's), instead of from small to large.

Individual Sales for 100 Salesmen

50	124	179	225	121	133	141	153
275	67	71	83	163	176	187	194
90	96	105	114	133	156	177	215
125	176	155	165	222	231	243	254
135	184	192	213	265	272	285	297
230	251	276	298	251	314	340	370
250	76	91	126	112	133	156	176
130	137	144	147	197	214	224	233
150	157	167	165	245	254	267	276
177	183	193	195	288	297	245	261
194	210	223	231	311	330	351	383
244	254	276	294	213	276	299	354
296	313	96	114				

11. The following summary presents certain information concerning some 90 employees of a banking organization. Each set of paired items represents an individual employee. The management administers an entrance test to each new employee, and the test scores refer to performance on this test. Throughout the period of employment, employees are rated semi-annually by their superiors, and the ratings were so obtained.

Perform the following operations in the preliminary statistical analysis of these data:

(a) Rearrange each set of scores (tests and ratings) so that it forms an array.

(b) Note the range of each set of scores.

(c) Prepare a frequency distribution of each set of scores.

(d) Prepare a double-frequency distribution of test scores (X) and ratings (Y).

TEST SCORES AND RATINGS OF 90 EMPLOYEES

Test score	Rating	Test score	Rating	Test score	Rating	Test score	Rating
78	94	94	94	81	70	94	85
67	80	98	97	89	82	91	82
75	82	100	100	97	91	81	67
76	70	79	82	72	75	82	73
92	86	75	77	79	89	77	76
66	74	82	86	84	87	99	90
99	98	98	93	94	92	84	77
99	83	79	81	93	89	80	78
78	78	85	86	81	68	91	83
94	79	81	77	86	93	81	81
58	78	82	81	71	77	93	87
85	75	74	72	87	89	77	79
88	94	89	93	71	84	83	96
80	91	74	81	88	92	74	74
80	73	93	88	75	78	87	83
92	82	97	93	91	90	95	93
82	75	75	83	79	76	79	76
79	77	75	74	79	78	80	82
98	81	91	95	92	84	60	65
71	86	95	97	77	79	65	68
70	56	77	79	91	93	90	88
83	89	62	65	86	82		
88	94	100	91	65	74		

CHAPTER IV

GRAPHIC REPRESENTATION

The meaning of statistical data may frequently be made more apparent by some form of graph or chart than by reference to detailed figures. For that reason various means of graphic representation are widely used, and the study of modern statis-
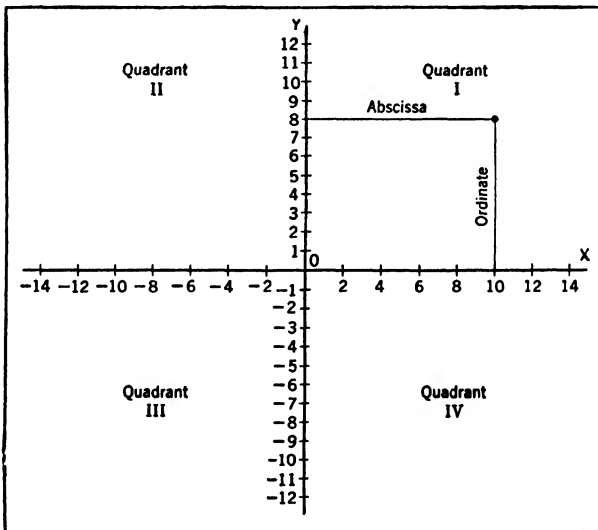
FIG. 4·1.—Rectangular Coordinates.

tical techniques must include consideration of the most important types of graphs.

Basic to most statistical charts is a conception of two-dimensional variation such as that pictured in Fig. 4·1. Such a graph makes reference to two rectangular coordinates, designated in the figure as X, the horizontal line, and Y, the vertical axis. The point of intersection of the two lines, called the *origin*, is designated O or R. It is customary to define scales on each of the axes, as has been done in the figure. The scales

48

begin with O at the origin and increase upward along the Y axis and to the right along the X axis.   They become negative below the point of origin on the Y axis, and they are negative also to the left of the origin on the X axis.

The two intersecting axes thus create four quadrants which are generally designated as shown by the Roman numerals in the figure.   The upper right-hand quadrant, which is most commonly used, is quadrant I.   It is, as will be noted, a positive quadrant, i.e., the values of both X and Y for this quadrant are positive.   Other quadrants are numbered II, III, and IV, in a counterclockwise rotation.

When the two lines are scaled, any point in one of these quadrants may be readily located by reference to these scales. Thus, in the figure, the point $X = 10$, $Y = 8$ has been marked by a small dot.   Its location is determined first by reference to the X axis and then to the Y axis.   The distance from O to 10 measured along the X axis is designated the *abscissa* of the point; the distance along the Y axis from O to 8 is called the ordinate of the point.   It will be apparent that given pairs of X and Y values can always be represented by an appropriate point in such a chart, and that, conversely, any point in any quadrant can be described in terms of its appropriate X and Y values.   In abbreviated form, such a point is frequently described by two numbers separated by a comma, as *10, 8,* which implies an X value of 10 and a Y value of 8.   The area included within the coordinate rulings upon which data are charted is called the *grid*.   Many standard grids are available, and special forms may be prepared for unusual graphs.

Dependent and independent variables.—The X and Y scales may be used to represent an almost limitless number of different kinds of data. X, for instance, may represent time, the passage of years (probably its most common usage in such charts), but it may also represent numbers of dollars, other measures of value, of weight, of size, or of other characteristics.   Only one rule can be stated as a limitation on the use of these scales: generally the variable regarded as *independent* is measured on the X axis, and the *dependent* variable on the Y axis.   There is, however, no entirely satisfactory test of dependence.   In

general, if a known logical cause-and-effect relationship exists, then the cause is regarded as independent and the effect as dependent. Thus, if rainfall and crop yields are compared, rainfall would be regarded as the independent variable; or, in a comparison of temperature and the changing viscosity of tar, the first would be considered the independent variable. Time is generally regarded as an independent variable in comparisons of time-to-time changes, since it can scarcely be regarded as a result of the changes. On the other hand, there are many situations in which this relationship is not clear and in which either variable might be regarded as independent. This is true, for instance, in comparisons of height and weight of human beings, or in similar relationships.

Functional relationships.—Few statistical charts will be confined to the location of a point or even of several points. In many, straight and curved lines will be used, for which reason it may be well to note the fact that such lines represent simple mathematical *functions*. Hence, if the mathematical function is known, the line may be located and used to represent it.

Such functions may be described in the same symbols as those that have been used to designate the two axes. Thus, if data are so related that Y always equals X, then location of two or more points will clearly indicate the fact that the function appears graphically as a straight line passing through the origin, O (or R). When $X = 1$, $Y = 1$, and when $X = -1$, $Y = -1$, and the representation of this function is the line shown as MN in Fig. 4·2. Again, if the relationship is $Y = 2X$, then, when $X = 1$, $Y = 2$; when $X = -1$, $Y = -2$; and a straight line such as PQ in Fig. 4·2 is the appropriate graphic representation. It will be noted that every point in such a line represents X and Y values such that $Y = 2X$.

Any straight line may, in turn, be represented by an equation, and that to which reference is most common defines the straight line as $T = a + bX$. In this equation, T refers to the Y value of a point in the line which accompanies a given X value.[1] The a in the equation refers to the Y value of the

---

[1] The T here used is an abbreviation for "trend," which is explained in Chapter X. Other symbols frequently used for this purpose include Y' and Y_c (computed Y).

line where it crosses the Y axis, and is sometimes called the
Y *intercept*. The b in the equation is a measure of *slope* in the
line; it refers to the amount of Y value that is added with a unit
increment of X. Thus, in effect, the equation says that the
height of the line (measured on the Y scale) for any given value
of X is the height at $X = 0$ (the value of a) plus X times the
slope. Hence, if values of a and b are known, the line may be
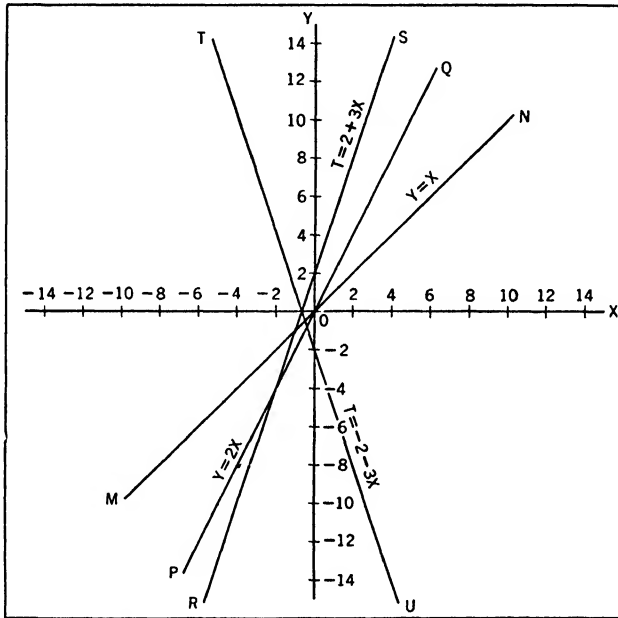
Fig. 4·2.—Linear Functions.

readily located. In many statistical problems, the major task
is that of discovering the appropriate values of a and b. These
values, it may be noted, can be either positive or negative. If
a is negative, the line crosses the Y axis below the point of
origin. If b is negative, the line slopes downward to the right
instead of upward. Thus, if $a = 2$ and $b = 3$, then, when
$X = 0$, $T = 2$ (the Y intercept). When $X = 4$, $T = 2 +$
$(4 \times 3) = 14$. The line thus defined is shown in Fig. 4·2 as RS.
Again, if $a = -2$ and $b = -3$, the line appears as TU in the
figure.

On many occasions, of course, straight lines are not adequate to represent the data under consideration. Resort may then be had to various types of curves. In many of them, the functional relationship of X and Y may be fairly simply stated, but others require more complicated representation. Reference will be made to some of these curves in the discussion of trend fitting in a later chapter.

Line charts; time series.—Probably the most common of all statistical charts is that which portrays changes in data over a period of time. For that purpose, the simplest and most commonly used chart is that in which a line connects points representing monthly, quarterly, or annual data. This type of chart
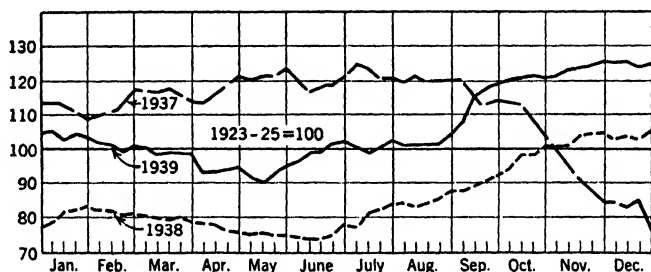
FIG. 4·3.—Line Chart: Time Series. Data: *Business Week's* index of business activity from *Business Week*, January 6, 1940, p. 11, by permission of the publishers.

is illustrated in Fig. 4·3. In such a chart, time, whether designated in days, weeks, months, quarters, years, or longer periods, is measured on the X axis. Either vertical lines or the spaces between such lines may be designated in time units. Where data represent averages of weeks, months, or years, it is probably preferable to represent them in the middle of spaces, rather than on one of the lines. The Y scale is a simple one, and the value for each week, in this case the index of business activity (the meaning of such an index is explained in Chapter VIII), is represented by a point. To aid the eye in following the series of movements, these points have been connected by straight lines.

No great difficulties present themselves in the construction of such a chart. Variations in style are common. What is important is that the title be accurate; that the scales and any

labels be clearly legible; that lettering be so positioned that the chart is easily read in its normal position or, if necessary, by turning it one-quarter turn clockwise; and that the Y scale be effectively labeled. If it is impractical to show the zero line on the Y scale, then that scale should usually be broken, as shown in Fig. 4·4, to call attention to this fact. In certain types of chart, notably those in which logarithmic scales are used, it is, of course, impossible to show the zero line.
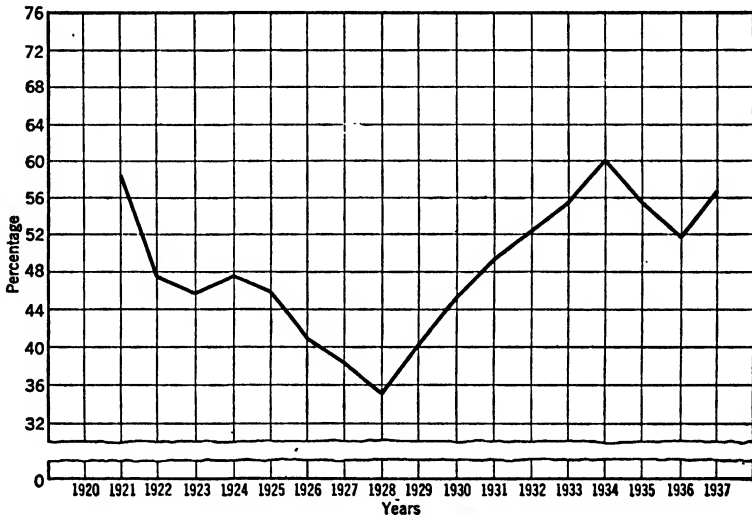
Fig. 4·4.—Illustration of Broken Vertical or Y Scale. Data: Percentage of families provided with one-family dwellings, from "Statistics of Building Construction, 1920–1937," Washington, Bureau of Labor Statistics, 1938.

Sometimes time-to-time changes in several different series may be represented in a single chart. For this purpose, different kinds of lines should be drawn, particularly if the lines cross, as is indicated in Fig. 4·5. The most commonly used of such lines are generally described as the solid line, the dotted line, the dash line, and the dot-dash line. Other combinations may be used, but in order to avoid confusion, it is usually undesirable to include more than four or five types of lines on the same chart.

The drawing of cross lines through the body of the chart, representing the divisions of X and Y scales, has become largely a matter of judgment and preference. Usually, however, a few

of them are drawn in at regular intervals or at important points. Thus a vertical percentage scale may be represented at 10-point intervals, and the 100 per cent line might be drawn somewhat heavier than the others. In other charts, 5-point intervals might be more convenient. Vertical lines may be drawn representing each month, each year, or each 5- or 10-year period. A chart is clearer if the cross lines representing the scale are rather far apart and are drawn more lightly than the lines and
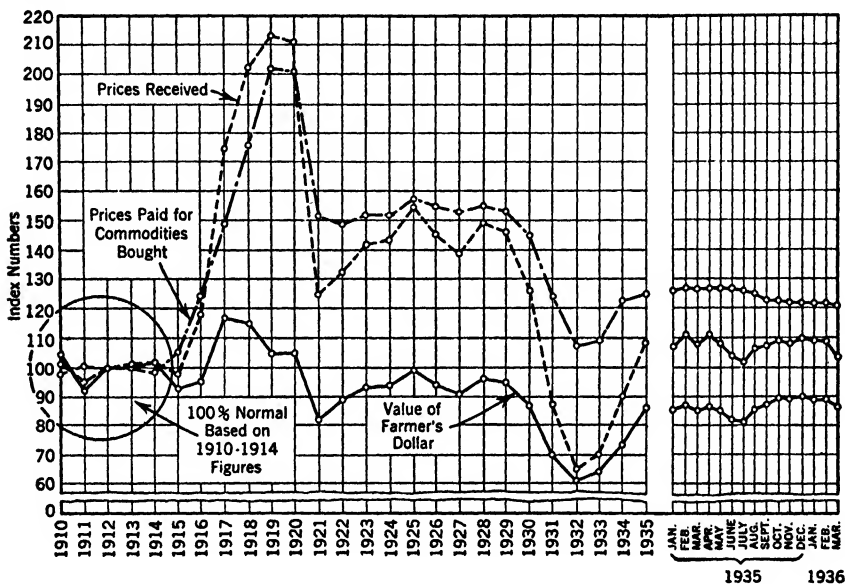
Fig. 4·5.—Multiple Line and Broken Time-Scale Chart. Data: Farm prices in the United States, 1910–1936. Source: United States Department of Agriculture.

points representing the data. Sometimes the Y scale, usually represented at the left, is repeated at the right of the chart, and this procedure is especially desirable if no cross lines are drawn.

Charts of frequency distributions.—In the preceding chapter, attention has been given to the nature and usefulness of frequency distributions. Such data may be portrayed graphically, and several types of charts are commonly used for this purpose. The most common is the *histogram* or *column* diagram, illustrated in Fig. 4·6. The data, representing daily sales of 40 employees in dollars, have been tabulated in class intervals of $5 each.

Since there were no sales below the \$15–\$20 class, that portion of the chart might have been omitted. Numbers of salesmen in each class are measured on the Y scale, and the size of each class is represented by a rectangle. The base of each rectangle represents the measure of the class interval and is bounded by the upper and lower class limits. Each class is thus portrayed as an area proportional to the total sales of all employees, which is represented by the total area of all the rectangles.
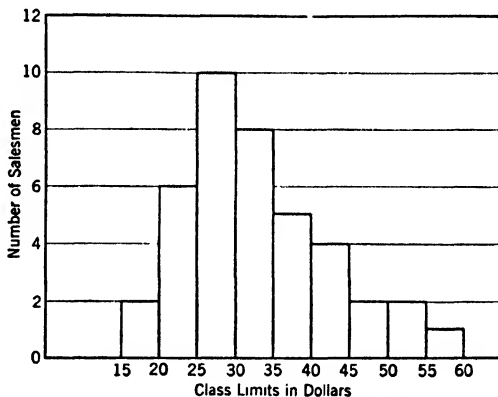
Fig. 4·6.—Histogram or Column Diagram. Data: Sales of 40 retail salesmen.

The usefulness and effectiveness of such a chart depend largely on the selection of class intervals. Too large an interval may obscure significant variations within the classes; too small an interval may result in a chart that fails to portray the general character of the data and overemphasizes peculiarities of individual items and classes.

The frequency polygon.—A frequency distribution may also be represented by what is generally described as a *frequency polygon*, illustrated in Fig. 4·7. Class limits are designated on the X axis, as in the histogram, but frequencies in each class, measured on the Y scale, are designated by a point directly

Fig. 4·7.—Frequency Polygon. Data: Sales of 40 retail salesmen.

above the mid-point of each class. The points thus located are then connected with lines, and the figure is usually closed at each end by bringing lines down from the mid-points of the outside classes, as plotted, to the mid-points of hypothetically adjacent classes on the base line, except in the case of an open-end class as noted later in this discussion. Such a chart is particularly useful where two or more frequency distributions are to be plotted on the same chart, in which event the polygon stands out more clearly than the rectangular form. The frequency polygon illustrated in Fig. 4·7 represents the same data as those portrayed in Fig. 4·6.

When the class intervals of a frequency distribution are of varying size, the charting, in either the rectangular or the polygon form, becomes somewhat more complicated. Suppose, for example, that most of the class intervals of a frequency distribution are 100, but there is also an interval of 200 and one of 500. The frequencies of the classes having increased intervals will obviously be increased over what they would have been had the intervals been uniform. Hence the frequency of the doubled interval should be plotted at one-half its given height; and the frequency of the fivefold interval should be plotted at one-fifth its given height. In effect, this makes the doubled interval into 2 classes, and the next interval becomes 5 classes. The general rule in such cases may be stated as follows: calculate the ratio of the standard interval to the given interval, and multiply the given frequency by the ratio thus obtained. The result is the frequency to be plotted, (f_p). That is, the frequency to be plotted is

$$f_p = f \times \frac{i_s}{i_g}$$

when i_s is the standard interval and i_g is the given interval. If the interval is standard, this formula indicates no change. If there is an open class at one or the other of the extremes of the distribution, the appropriate height may be estimated. Sometimes this estimate is plotted as a horizontal dotted line of indefinite length or a point at an assumed class mark, and sometimes it is omitted entirely. Usually the frequency in such a class is comparatively small.

It should be noted that the frequency polygon may become almost a smoothed frequency curve if data and classes are numerous. As class intervals are reduced, under such circumstances, the smoothing process is increased, and if the data are characterized by no striking variations a smoothed curve replaces the regular polygon. The data used in this discussion are not sufficiently numerous to provide an illustration of this process, but Fig. 4·8 has been constructed from a much more extensive
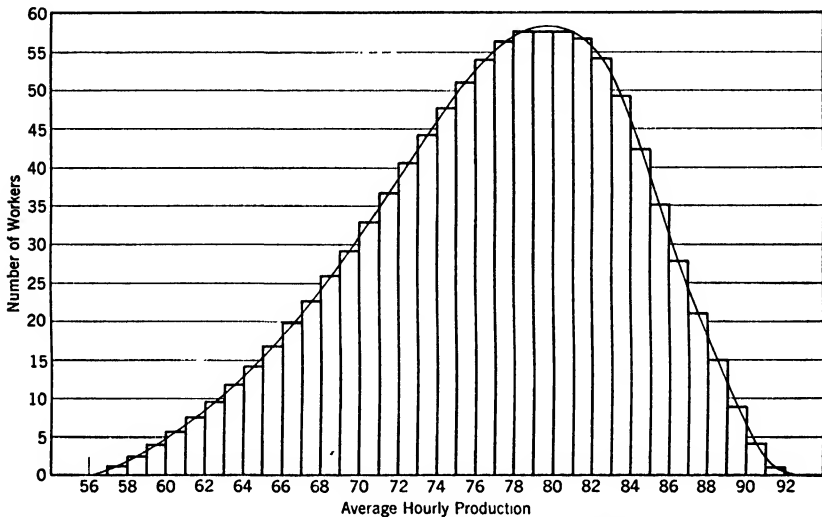
Fig. 4·8.—Derivation of Frequency Curve from Column Diagram. Data: Numbers of employees classified according to average hourly production. Source: confidential.

tabulation to show how such a distribution may be represented by a frequency curve.

Cumulative frequency charts.—For many purposes, the most convenient chart of a frequency distribution is one that represents its cumulative frequencies. These frequencies are plotted against their respective limits: Σ_1 against L_1, Σ_2 against L_2, etc., and the limits are commonly shown on the X scale, cumulative frequencies being measured on the Y scale. Such a chart, commonly described as an *ogive*, is pictured in Fig. 4·9, which plots the cumulatives of the distribution shown in Figs. 4·6 and 4·7. In one of its two forms, illustrated by Fig. 4·9, the curve begins with a zero frequency at the lower limit of the first

class and depicts the total frequencies "less than" each upper limit. Sometimes a similar curve is drawn, using the reversed summations or cumulatives to represent total frequencies "more than" each class limit. Such a chart, for the same data, is illustrated in Fig. 4·10. In practice, both types of cumulatives are generally represented by smoothed curves rather than by short straight lines between class limits. The smoothing achieves a result similar to that which would appear if the class interval were reduced and the number of classes were greatly increased. The ogive is widely used and is a particularly effective form of graphic representation. Sometimes cumulative percentages are so represented, instead of cumulative frequencies.

FIG. 4·9.—"Less Than" Cumulative Frequency Curve. Data: Sales of 40 retail salesmen.

The Z chart. — Another method of representing cumulative data and comparing it with other series that has come into increasing use in recent years is the Z chart, so called because its lines form a more or less crude Z. In its most common form, the Z chart compares individual months with totals for the year and with a *moving total* for the 12-month period ending in each month. This type of chart may best be explained by reference to specific data. Those summarized in Table 4·1 will serve this purpose. It will be noted that there are three items for
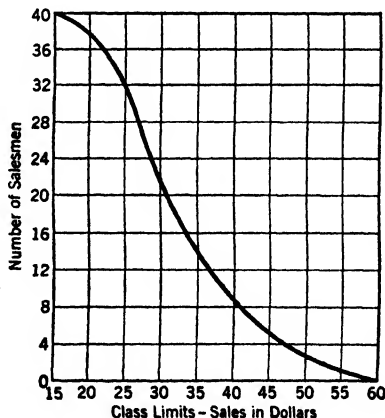
FIG. 4·10.—"More Than" Cumulative Frequency Chart. Data: Sales of 40 retail salesmen.

each month: (1) the sales in that month; (2) the cumulative total of sales in the year; and (3) the total sales for the 12-month period ending in the given month. Each of these items is represented in the chart, which is illustrated by Fig. 4·11. Monthly

TABLE 4·1

MAIL-ORDER AND STORE SALES OF SEARS, ROEBUCK AND CO., 1938
(Thousands of Dollars)

Source: *Survey of Current Business*, 19 (3), March, 1939, page 25.

Month	Monthly sales	Cumulative sales	12-month moving total
January	30.6	30.6	572.7
February	30.5	61.1	571.4
March	41.1	102.2	568.8
April	44.9	147.1	564.1
May	43.5	190.6	554.1
June	43.8	234.4	545.7
July	36.3	270.7	538.8
August	39.9	310.6	537.1
September	49.2	359.8	533.5
October	53.3	413.1	528.2
November	51.2	464.3	529.1
December	68.6	532.9	532.9
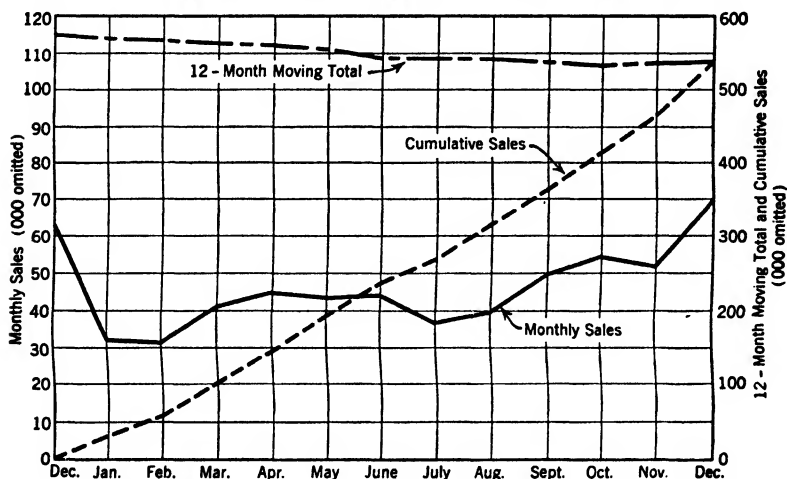
FIG. 4·11.—The Z Chart. Data: Retail and mail-order sales of Sears, Roebuck and Co., as shown in Table 4·1.

sales are shown by a solid line; the cumulative is a dotted line; and the moving total is a dot-dash line. It may be noted that the cumulative sales figures are readily obtainable from the data of the table, being simply the totals of given monthly figures. Thus the cumulative for February is simply the total of January and February. The moving 12-month totals, except for December, cannot be calculated from the data of the table, since they require reference to the preceding year. All the items for each month are customarily plotted on the ordinate representing that month rather than between these ordinates.
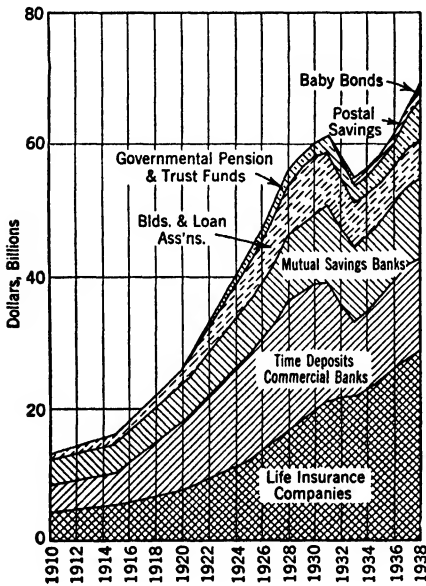
FIG. 4·12.—Component Area Chart. Data: Growth of savings institutions in the United States, 1910–1938. Source: *Business Week*, May 27, 1939, p. 18.

Component parts.— Frequently, a variation of the ordinary line diagram may be used to show the component parts that make up a given series. Thus, Fig. 4·12 represents the growth of savings institutions in the United States. It indicates, also, the totals of the seven types of securities that make up these investments. Such a chart makes apparent not only the changes in total savings but also the shifts in such savings from year to year.

Bar charts.—Sometimes, when the number of periods or items to be compared is not extensive, statistical data may be effectively portrayed by means of bar charts. A simple type of bar chart has already been illustrated in the discussion of the column diagram as a portrayal of the frequency distribution. There, it was noted that rectangles were erected on a base representing the class interval and to a height which measures the frequencies on the Y scale. In most cases, the bar chart con-

sists of bars of uniform width separated by spaces about half
that width. Bars may be either vertical or horizontal, but com-
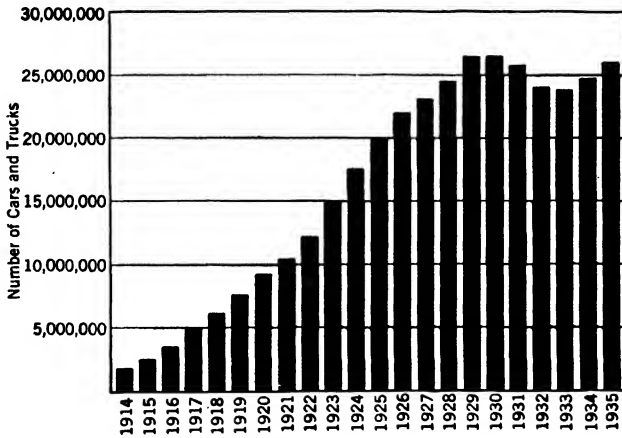
FIG. 4·13.—Simple Bar Chart (Time Series). Total registration of cars and trucks
in the United States, annually, 1914–1935. Source: 27th Annual Report of General
Motors Corporation, p. 51.
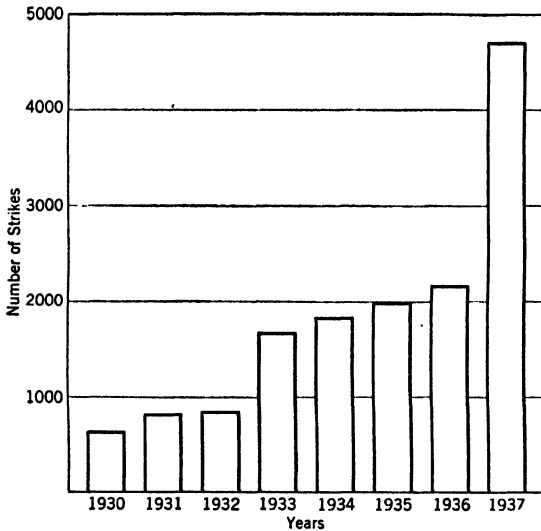
FIG. 4·14.—Outline Bar Chart. Data: Annual numbers of strikes in the United
States, 1930–1937. Source: Bureau of Labor Statistics.

mon practice uses vertical bars for time series. For other data,
horizontal bars are frequently more convenient, because they
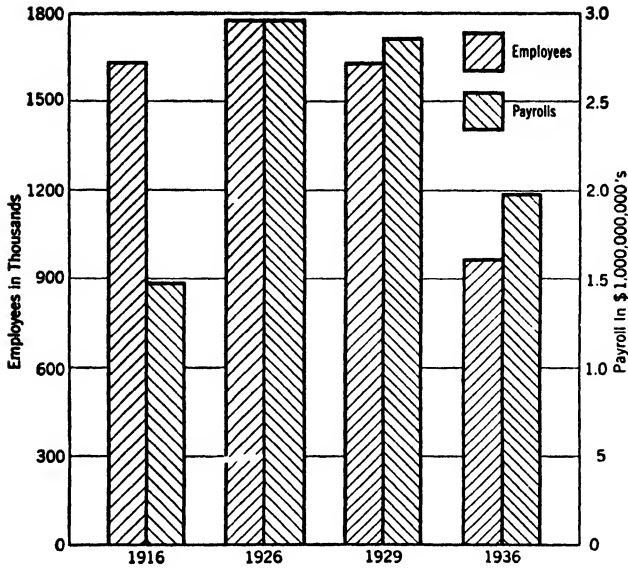
FIG. 4·15.—Comparative Bar Chart. Data: Employment and payrolls on Class 1 railways, selected years. Source: *A Yearbook of Railroad Information*, 1937 Edition, p. 60.
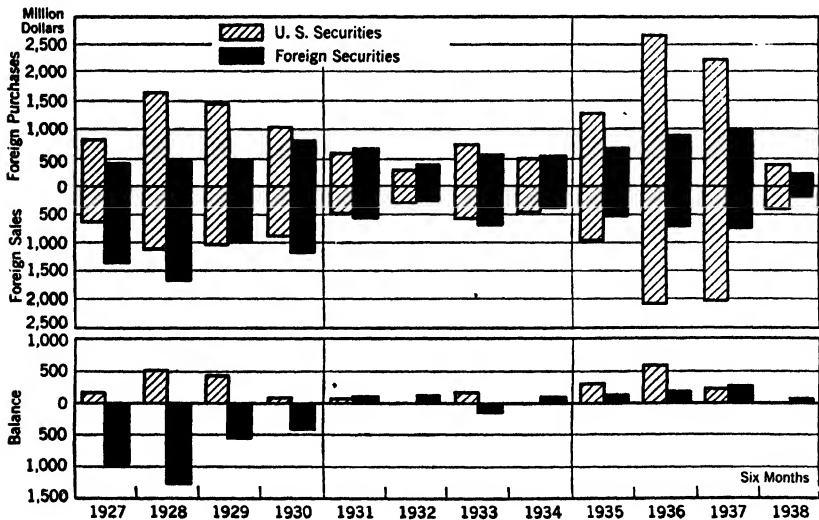
FIG. 4·16.—Bar Chart with Positive and Negative Scales. Data: Capital transfers between United States and foreign countries resulting from security transactions. Source: *New York Stock Exchange Bulletin*, Vol. IX, No. 10, October, 1938, p. 1.

allow more space for labeling. The scales involved are usually placed along the left-hand margin and at the top or bottom. The comparisons thus permitted are, of course, based on the length of the individual bars, since the width of bars is the same. Bars may be shown as solid or merely in outline. Simple bar charts of both types are illustrated in Figs. 4·13 and 4·14.

Sometimes, it is convenient to use two or more bars, distinguished by shading or hatching, in contrast with each other. Such a chart is shown in Fig. 4·15. Again, bars may be arranged on both sides of a zero line or origin, thus portraying positive and negative values. Such usage is illustrated in Fig. 4·16.

Component bar charts.—The bar chart is admirably adapted to the purpose of showing component parts. In such cases, bars are divided, and different types of shading distinguish the components that make up the total represented by the area or height of the bar. Guide lines, shown in Fig. 4·17, help lead the eyes from one bar to another and suggest the nature of changes in the various components.

Pie charts.—One of the most commonly used graphic representations is the circle or pie chart. It is poorly adapted to some purposes, notably where comparisons between quantities are represented by varying areas of circles, but it is effective as a means of showing components. Figure 4·18 illustrates its use for this purpose.

Pictorial charts.—Recent years have witnessed an increasing use of various pictorial charts. In general, they seek to "dress
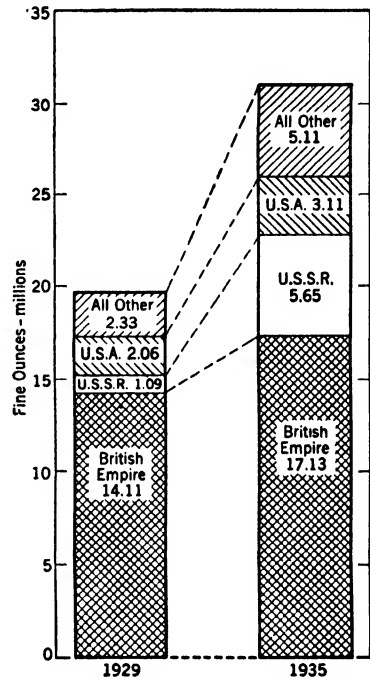
FIG. 4·17. — Component Bar Chart. Annual gold production of the world, 1929 and 1935. Source: *New York Stock Exchange Bulletin*, Vol. VII, No. 6, June, 1936, p. 1.

up" the more or less prosaic lines, bars, and circles and give them "eye appeal," thus attracting attention. Some of them are both effective and reasonably accurate; others are likely to prove misleading. If they seek merely to effect comparisons in one dimension, as, for instance, in the length of a pictorial bar, there can be little objection to them. When, however, they attempt to represent changes or differences in magnitude by differing areas or by three-dimensional portrayals of volume,
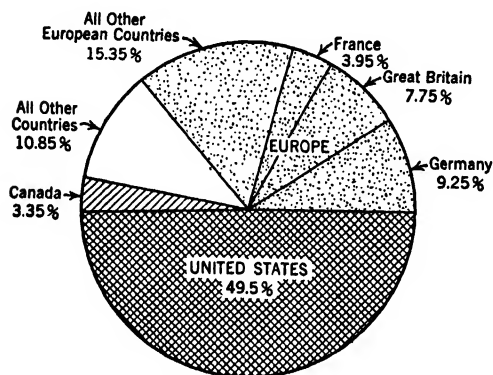
Fig. 4·18.—Pie Chart or Circle and Sector Chart. Data: World distribution of telephones, 1938. Data from American Telephone and Telegraph Co. as reported in the *World Almanac*, 1940, p. 544.

they are much less effective. Human observation appears generally unable to derive effective comparisons of volume from the usual graphic representations. Moreover, it is not infrequently impossible to tell, in charts where quantities are represented by single figures of varying size, whether comparisons are to be made on the basis of area, volume, or some other standard. Several of the more satisfactory types of pictorial charts, in which comparison is made on the basis of the length of bars, are presented in Figs. 4·19 and 4·20.

Statistical maps.—One of the most frequently useful types of graphic representation is the statistical map. Several types of such maps are illustrated in the figures of this chapter. Some of them represent comparatively small areas, such as counties, cities, trade centers, and similar units; others may be used to portray or compare conditions characteristic of states, nations, or more extensive areas.

In what is perhaps the simplest form of statistical map, items of interest are merely located with reference to some central point, and their location is designated by dots, circles, small squares, or other appropriate symbols. Figure 4·21, for

A Dollar's Worth of Bread

Feb. 15, 1929
11.4 One Pound Loaves

Feb. 15, 1933
15.4 One Pound Loaves

Feb. 12, 1935
12.0 One Pound Loaves

Feb. 11, 1936
11.9 One Pound Loaves

A Dollar's Worth of Milk

Feb. 15, 1929
6.9 Quarts

Feb. 15, 1933
10.3 Quarts

Feb. 12, 1935
8.5 Quarts

Feb. 11, 1936
8.5 Quarts

Labor Time Spent in Producing 100 Barrels of Cement

Fuel, Power, Etc., 20.6 Man-Hours

Quarrying and Manufacturing, 55.0 Man-Hours

Transportation to Site of Construction, 50.7 Man-Hours

Each Figure Represents 5 Men Working 1 Hour

Barrels of Cement Produced by 100 Man Hours-of Work

1925   164 Barrels

1929   211 Barrels

1934   242 Barrels

Each Symbol Represents 25 Barrels

Fig. 4·19.—Pictorial Bar Charts. Source: *Labor Information Bulletin,* February and April, 1936.

STEEL PRODUCTION

WEEK ENDING
JANUARY, 22
1938

JANUARY, 21
1939

Each symbol represents 50,000 tons of ingots

AUTOMOBILE PRODUCTION

WEEK ENDING
JANUARY, 22
1938

JANUARY, 21
1939

Each symbol represents 10,000 automobiles

PICTORIAL STATISTICS, INC.
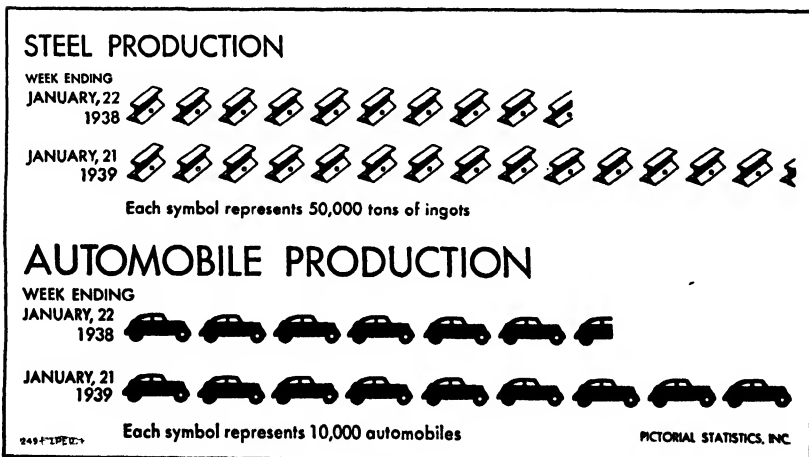
Fig. 4·20.—Pictorial Bar Charts. Source: *Monthly Survey of Business,* American Federation of Labor, No. 99, January, 1939.

instance, portrays the location of subscribers to a small city newspaper with reference to the city in which it is published.

LOCATION OF 519 RURAL WASECA NEWSPAPER SUBSCRIBERS
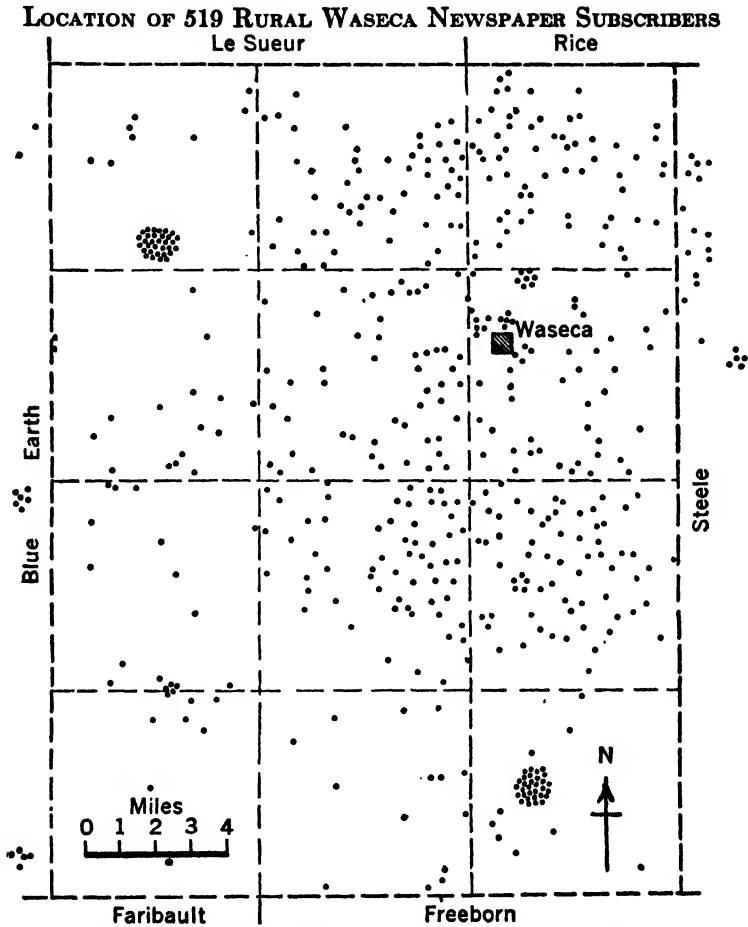
FIG. 4·21.—Dot Map. Data: Location of 519 rural Waseca newspaper subscribers. Reproduced by permission from Ralph Cassady, Jr., and Harry J. Ostlund, *The Retail Distribution Structure of the Small City*, Minneapolis, University of Minnesota Press, 1935, p. 31.

It effectively indicates the trade area served by this advertising medium.

The symbolic map, illustrated in Fig. 4·22, is less frequently used but is an effective representation. Here selected symbols

FIG. 4·22.—Symbolic Map. Data: Geographic distribution of women workers in five industries. Source: "Employed Women under N.R.A. Codes," *Women's Bureau Bulletin* 130, Washington, 1935.

represent the special characteristics to be distinguished and compared, and it is the purpose of the chart to be attractive as well as informative, to entice the reader to examine it and the situation it presents.

The two characteristics that are most commonly illustrated by statistical maps are density and quantity; a third less common use seeks to distinguish qualitative differences between various sections. In comparisons of density, most usage prefers the dot map, illustrated in Fig. 4·23. Generally, such a map is most effective when dots are of uniform size and are so arranged that their numbers in a given area are proportional

WHEAT PRODUCTION 1922
(U.S.D.A. Estimates)
Each ● represents 100,000 bushels
or major fraction thereof
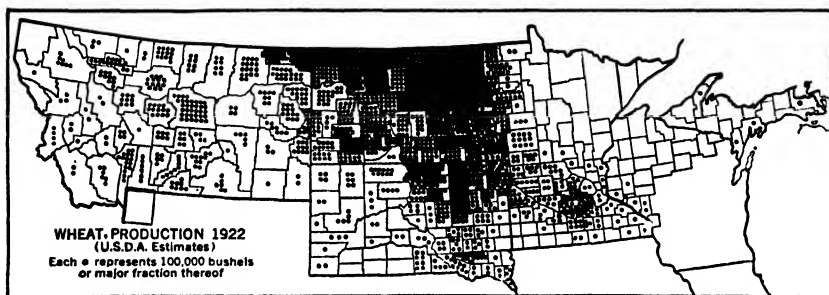
Fig. 4·23.—Dot Map. Quantitative Comparisons. Data: Wheat production in the United States, 1922. Source: *Graphic Survey of Agricultural and Financial Conditions in the Ninth Federal Reserve District*, p. 71, published by the Ninth Federal Reserve Bank, Minneapolis, 1923.

to the density of the section. Care must be exercised to prevent the dots from running together. Sometimes the effect of the dot map is secured by placing large-headed pins in a map that is properly reinforced with heavy cardboard or cork. Such pin maps are in general business use and are commonly associated with displays of marketing organizations. There is always the possibility of using various-colored pins to indicate variations in the data, and the pins may be readily moved about to show changing business conditions.

Where comparative size, magnitude, or quantity is to be illustrated, as in comparisons of crop yields, mineral resources, output of manufactured products, consumption habits, or similar data, maps may be variously cross-hatched, or solid bars or

circles may be drawn within the various areas, thus providing an effect similar to that of the dot map of density. For this reason, the cross-hatching is generally preferable.

Figure 4·24 illustrates a statistical map in which cross-hatching is used to present qualitative rather than quantitative differences. Sometimes, in such maps, the shading is arranged so that it is heaviest where the condition under consideration is most prominent, so that, in a sense, it attempts a quantitative

Fig. 4·24.—Statistical Map. Qualitative Comparisons. Data: Minimum wage laws for women. Source: "Summary of State Hour Laws for Women and Minimum Wage Rates," *Women's Bureau Bulletin* 137, Washington, 1936, p. 10.

presentation. Thus, lighter-shaded areas have less of the given characteristic, while black areas represent the sections where the opposite condition prevails.

The nomograph.—A very special type of chart, called a *nomograph*, which may be prepared to facilitate various types of computation, is illustrated in Fig. 4·25. This chart is so scaled that if two numbers are located on scales *A* and *C*, respectively, a ruler connecting them will locate on scale *B* the product of the two numbers. If the two numbers are alike, obviously the square will be obtained. Conversely, division may be carried out by noting the product on *B* and one factor on *A*,

FIG. 4·25.—Nomograph for Products and Squares. Courtesy of the Codex Book Co., Inc., Norwood, Mass.

and, by means of a ruler, locating the corresponding factor on C. Or a square root may be obtained by locating the square on B and, holding the ruler in a horizontal position, reading the square root on A and C. Decimal points are determined by inspection. This nomograph is shown merely as an illustration of a type of chart which is adaptable to a wide range of uses. In laboratory practice, nomographs are frequently prepared to facilitate commonly encountered types of calculation. For a discussion of the subject and a presentation of many useful nomographs, the student is referred to the *Handbook of Statistical Nomographs, Tables, and Formulas*, by Dunlap and Kurtz.[1]

The Lorenz curve.—Another method of portraying certain types of comparisons is illustrated by the widely used *Lorenz curve*, which was devised by Dr. Max Lorenz, statistician of the Interstate Commerce Commission, to portray the distribution of wealth and income. It tends to emphasize all departures from an even distribution. Data are classified as shown in Table 4·2, and the chart, illustrated by Fig. 4·26, measures

TABLE 4·2

CUMULATIVE DISTRIBUTION, NON-FARM FAMILIES, 1929 (United States)*

Income class (dollars)	Cumulative families (thousands)	Cumulative per cent of all families	Cumulative income (million dollars)	Cumulative per cent of all income
0– 500	650	3.0	−440	−0.6
500– 1,000	2,735	12.6	1,230	1.8
1,000– 1,500	7,484	34.5	7,192	10.3
1,500– 2,000	11,578	53.4	14,309	20.5
2,000– 3,000	16,156	74.5	25,416	36.3
3,000– 5,000	19,496	90.0	38,041	54.4
5,000–10,000	21,047	97.1	48,395	69.2
10,000–50,000	21,611	99.7	58,518	83.7
50,000 and over	21,674	100.0	69,922	100.0

* Adapted from Maurice Leven, Harold G. Moulton, and Clark Warburton, *America's Capacity to Consume*, Washington, The Brookings Institution, 1934, p. 231, by permission.

[1] Jack W. Dunlap and Albert K. Kurtz, Yonkers-on-Hudson, N. Y., World Book Co., 1932.

cumulative percentages of the two distributions, as indicated by the designations of X and Y axes. If the distributions are similar, in that cumulative percentages parallel each other throughout the various classes, the chart tends to approximate a straight line. If there are disparities, as is true of the data
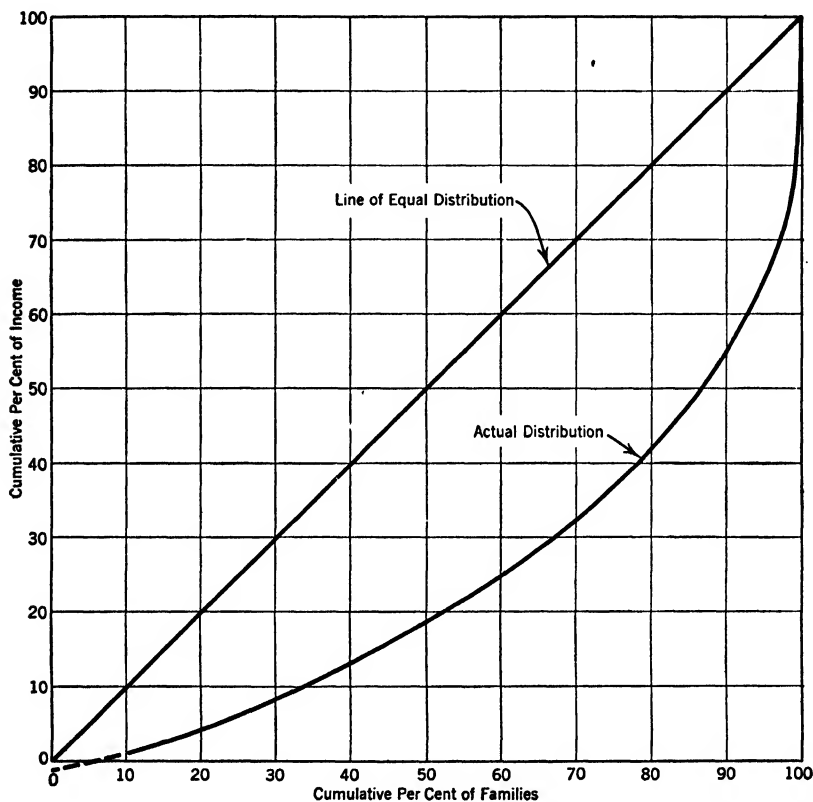
FIG. 4·26.—The Lorenz Curve. Data: Cumulative distribution of non-farm families by income groups, 1929. Source: See Table 4·2.

used here, that fact is apparent in the curvature of the line representing the actual distributions.

Logarithmic scales.—Thus far in this discussion of graphic representation, reference has been made only to charts that utilize arithmetic scales. In other words, the Y scale in a time series chart such as that shown in Fig. 4·4 or in a graph of a

frequency distribution such as Figs. 4·6, 4·7, 4·8, and 4·9 begins at zero and increases in the order of 1, 2, 3, 4, etc., by adding equal increments. The measures 1 and 2 are separated exactly the same distance on the chart as are 2 and 3 or 13 and 14. Similarly, the distance between 2 and 4 on the scale is twice as great as that between 1 and 2, but it is only half as great as that between 4 and 8.

There are, however, many situations in which the *rate of change* is of greater importance than the amount of change. It may be more significant that sales, for instance, increase by 20 per cent over August in September than that they increase by $2,000,000. This type of comparison may be particularly useful if such a September gain in one year is being compared with similar gains in other years. This is only a way of saying that the *ratio* of September to August or of the September gain to the August figure may be more important than the dollar increase.

Numerous other illustrations might be cited to illustrate the frequent importance of ratios. Thus a life-insurance agency compares its last year's addition to the total insurance in force. It may have added $5,000,000. But it has been growing each preceding year, and it is more interested in knowing whether it has continued its established *rate of growth* than in the actual amount of new business written. It wants to know how the *ratio* of new business to that on the books compares to the ratios of earlier years.

Where the *rate of change* or the *ratio* is the significant consideration, it is customary to supplant the ordinary arithmetic scale with a ratio or logarithmic scale. Such scales, illustrated in Fig. 4·27, allot equal space to equal ratios rather than to equal increments. Thus, there is the same spread from 1 to 2 as from 2 to 4, 4 to 8, 10 to 20, 100 to 200. Therefore, if the base scale on the X axis represents years in a simple arithmetic sequence, and the *ratio* scale is used on the Y axis, an increase of 20 per cent will result in a line having a given slope, no matter how much the base on which the 20 per cent is calculated may be. If, therefore, a firm increases its business 20 per cent one year and has a 20 per cent increase the next year (using the first

year as a new base), the line of its sales will be straight. If the second year's increase is less than 20 per cent, the slope is less than before. If it is more than 20 per cent, the slope is greater. An increase of exactly the same *amount* the second year would cause the line to show less slope, because the ratio would be smaller.

On the ratio scale, an arithmetic progression (i.e., one which increases by constant increments such as 1, 2, 3, 4, etc.) thus appears as a curved line, as shown in Fig. 4·28, while such a progression effects a straight line on an ordinary arithmetic scale. Conversely, on the ratio scale, a geometric progression (one which increases by constant proportions, such as 1, 2, 4, 8, etc.) appears as a straight line, while a geometric progression charted on an arithmetic scale curves rapidly upward, as shown in the figure.

To portray such changes in rates of change the ratio chart makes use of a *logarithmic* scale. That scale locates various numbers at the point represented by their logarithms rather than by the numbers themselves. Thus, since the logarithm of 1 is 0,·1 is placed on the base line. The logarithm of 2 is 0.3010, so 2 falls where 0.3010 would fall on an arithmetic scale. Similarly, since the logarithm of 4 is 0.6021, 4 is located at 0.6021 on the usual arithmetic scale, and 40, whose logarithm is 1.6021, is located in the same manner. Exactly the same result could be achieved if, using an arithmetic scale, each number was plotted as its logarithm. In other words, a ratio chart might be constructed

Fig. 4·27.—Logarithmic or Ratio Scales.

using an arithmetic scale if, instead of a value such as 4 being located at the point indicated on the scale, it was plotted at a point representing its logarithm, 0.6021, and all other values were similarly plotted.

Logarithms.—For those whose recollection of logarithms is hazy, a brief summary of the most significant facts may be helpful. In the most commonly used system, the logarithm of a number simply expresses that number as a power of 10. Thus 1 is the 0 power, and its logarithm is 0; 10 is the first power or
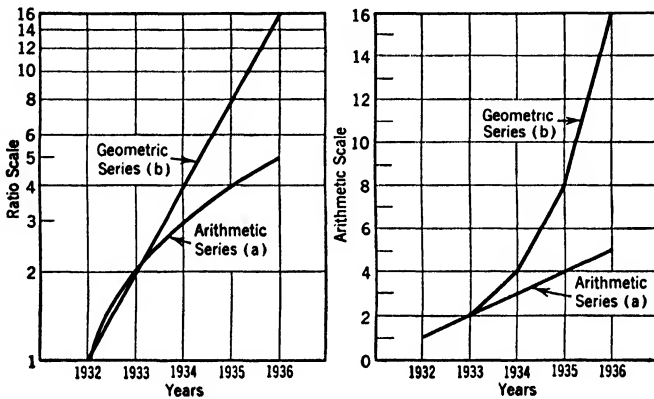
FIG. 4·28.—Comparison of Arithmetic and Ratio Scales. The series plotted are: (a) 1, 2, 3, 4, 5; and (b) 1, 2, 4, 8, 16.

10^1, and its logarithm is 1; 100 is 10×10 or 10^2, and its logarithm is 2; 1,000 is $10 \times 10 \times 10$ or 10^3, and its logarithm is 3. Most numbers are, obviously, not integral powers of 10. The number 15, for instance, is 10 raised to the power *1.1761* or $10^{1.1761}$; hence its logarithm is 1.1761.

Logarithms are frequently of great convenience in statistical analysis, for they often provide a means of shortening what would otherwise be long, tedious manipulations. Their convenience arises largely out of certain of their simplest characteristics, of which the most important may be summarized as follows: (1) numbers may be multiplied by adding their logarithms; (2) one number may be divided by another by subtracting the logarithm of the second from that of the first; (3) a number may be raised to any power by multiplying its logarithm

by the index of that power; (4) any root of a number may be taken by dividing its logarithm by the index of that root.

Before these procedures are illustrated, it may be well to note the usual rules for using tables of logarithms. The logarithm of any number is made up of two parts: (1) the integer before the decimal point, and (2) the decimal fraction that follows. Thus the logarithm of 20 is 1.3010. The integer, known as the *characteristic*, is 1, and the decimal, known as the *mantissa*, is .3010. The *characteristic* is readily determined without refence to any table, for it is the number of digits in the whole number less 1. Thus, for 1, it is 0; for 11, it is 1; for 75, it is 1; for 257, it is 2; and for 4,756, it is 3. If the number whose logarithm is to be taken is itself a decimal fraction, a similar rule prevails and the *characteristic* becomes negative. Thus, for 0.312, since the first digit is next to the decimal point, the *characteristic* is −1; for 0.0312, it is −2; for 0.00312 it is −3. For convenience in manipulation, such negative characteristics are frequently expressed in a somewhat different form by adding 10 to the *characteristic* and subtracting it from the whole logarithm. Thus:

$$\log 0.312 \quad = 0.4942 - 1 = 9.4942 - 10$$
$$\log 0.0312 \quad = 0.4942 - 2 = 8.4942 - 10$$
$$\log 0.00312 = 0.4942 - 3 = 7.4942 - 10$$

The *mantissa* of the logarithm is read from a table such as appears in the Appendix. There are many styles of tables, and they carry the logarithms to from four to ten or more decimal places. Usually, such tables are accompanied by directions, but their use is readily illustrated.

Suppose, for instance, that it is desired to multiply 4.32 by 43.78. It is apparent that the *characteristic* of the first logarithm is 0, and that of the second is 1. The *mantissa* for 432 is available from the table as .6355; that for 4,378 is .6413. The two logarithms are thus 0.6355 and 1.6413. Their sum is 2.2768. To interpret this total, it is necessary to discover its *antilogarithm*. The *mantissa*, .2768, is checked in the table, where it is found to designate a number whose digits are 18915. The *characteristic*, 2, indicates, however, that there are three

integers before the decimal, so the product must be approximately 189.15. A similar process is followed in division. If, for instance, 4.32 is to be divided by 43.78, the logarithm of the second is subtracted from the first. The subtraction is accomplished as follows:

$$\begin{aligned}
\log\ 4.32 &= 0.6355 = 10.6355 - 10 \\
\log 43.78 &= 1.6413 = \underline{\ 1.6413} \\
& 8.9942 - 10 = 0.9942 - 2
\end{aligned}$$

The remainder is 0.9942 − 2. The *mantissa* indicates a value of 9867, while the negative *characteristic* indicates that the decimal point is followed by one 0. The quotient is thus .09867.

Other uses of logarithms may best be explained where they are required in connection with specific statistical manipulations.

The wide range of uses to which the ratio scale may be put needs little description. It is applicable wherever *rates of change* or *ratios* are to be presented. Generally, the scale is applied to only one axis, the arithmetic scale being preserved on the other. The figure thus provided is known as a semi-logarithmic chart, since only one of its two dimensions is measured by the ratio scale. This type of chart is particularly useful in portraying rates of growth and in forecasting. It also facilitates a comparison of fluctuations, since it portrays rates of decline as well as of advance, and it is readily adaptable to the presentation of changes measured in widely differing units, since it emphasizes relative rather than absolute changes. An illustration of one of these uses is presented in Fig. 4·29.

Special scales.—Numerous special scales on one or both axes may be useful in particular types of analysis. Where, for instance, reference is made to probability and chance, a scale indicating the theoretical expectancies of the *normal distribution* may have advantages. This and other special types of charts are illustrated in later chapters.

Gantt charts.—Numerous concerns have found a type of chart developed by H. L. Gantt to be of wide usefulness, especially in production and sales control. The Gantt charts represent a modification and adaptation of the bar chart. They are available in a number of forms designed to meet a variety of

needs in business.   Some of the charts are adapted to records of men, others to the performance of machines, and still others to more complicated processes involving both men and machines.

The essential feature of the Gantt chart is its comparison of individual week and cumulative weekly totals with standards
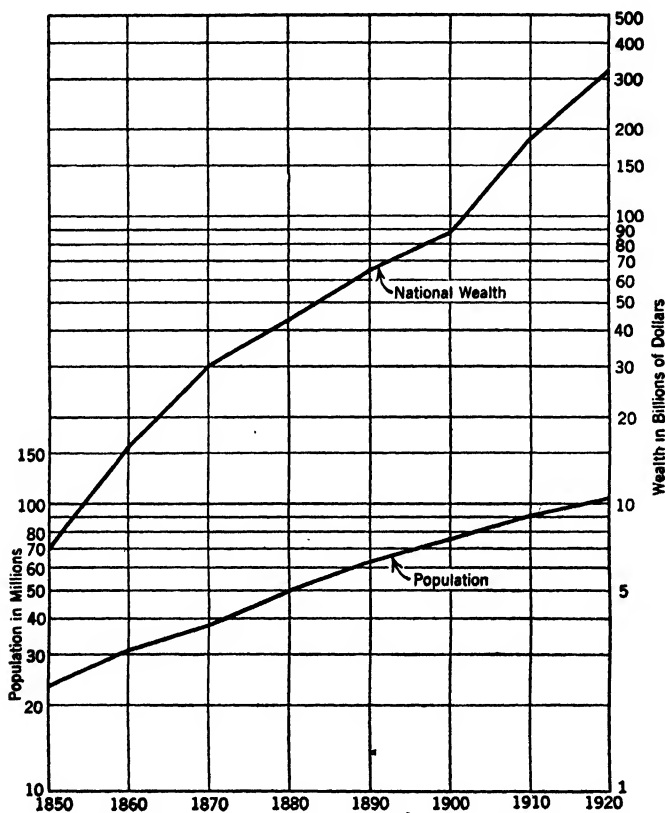
FIG. 4·29.—Use of Logarithmic Scale.   Data: Growth of population and of national wealth in the United States.   Source: *Statistical Abstract of the United States*, 1935, pp. 4 and 268.

or quotas assigned in advance.   In the typical chart, a heavy bar at the top, for instance, may represent the quota assigned. After each name on the chart, lighter lines then show individual weekly production as a percentage of the assigned quota, and the continuous bar following each name shows the cumulative

total for the period. The variety of applications to which such charts may be adapted is obvious.[1]

Charting techniques.—In conclusion, it may be well to summarize some of the more important general rules to be followed in preparing graphs. Primary among such rules is the requirement that charts be carefully, accurately, and completely labeled. Sources of the data should be described in such a manner that the reader can check on the accuracy of reporting. Scales must be as carefully designated, and care must be taken to select scales that portray the data and changes in them accurately and without misrepresentation. If scales are broken, i.e., if they do not follow a regular sequence or do not, in most cases, start at zero, that fact must be clearly indicated. All labels must be clearly legible and readable if the chart is held in its normal position or rotated one-quarter turn, clockwise.

The art of drawing presentable and effective charts cannot be adequately discussed in an elementary statistics textbook, but a few general features of modern charting procedure may be briefly described. Charts are generally first sketched in pencil on cross-section paper with scales appropriate to the data and the purposes for which the graph is intended. Such paper is readily obtainable from stationers and bookstores. These rough drawings are then attached to drawing boards and copied on tracing vellum or tracing cloth. The tracing is made in india ink with engineers' drafting pens, which may be conveniently adjusted to produce the required width of line. A rule with a steel edge to keep the ink from the paper is generally employed. Or, sometimes, the drawing on the original cross-section paper is inked without further copying. If the chart is to be reproduced as a cut for printing, it is generally drawn to dimensions considerably larger than those of the reproduction, and suitable allowance for the reduction should be made in the weight of the lines and the size of the letters.

Lettering is sometimes done freehand, but such a procedure generally requires considerable skill and training. Hence, it is commonly done by means of lettering guides, several varieties of which may be purchased. These guides come in sets of

[1] See Wallace Clark, *The Gantt Chart*, New York, Ronald Press Co., 1922.

various sizes and may be vertical, italic, and of varying design. Suitable stylus pens adapted to the guides are used. The guides are placed against a T square on the drafting board, and are shifted to the proper position on the chart. They may be suc-

Fig. 4·30.—The Stylus Pen Employed with Lettering Guides, by courtesy of the Wood-Regan Instrument Co., New York.
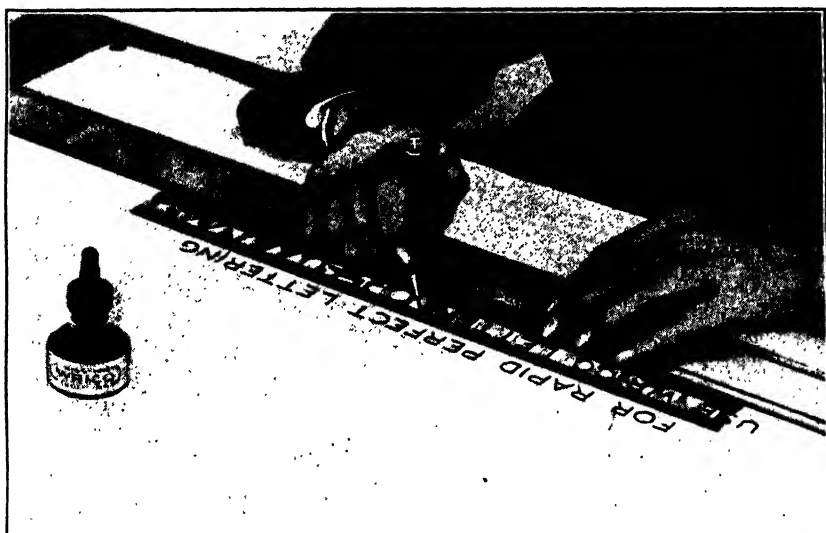
Fig. 4·31.—Use of Lettering Guides and Stencils, by courtesy of the Wood-Regan Instrument Co., New York.

cessfully used by anyone after a little practice. The accompanying pictures (Figs. 4·30 and 4·31) indicate the way in which they are used. Detailed instructions, of course, are furnished by the makers of these instruments.

READINGS

See "Classified readings from readily available texts," pages 591–597, also:

ABEL, JAMES F., "A Graphic Presentation of Statistics of Illiteracy by Age Groups," U. S. Office of Education, Pamphlet 12, April, 1930.

ARKIN and COLTON, *Graphs: How to Make and Use Them*, New York, Harper and Brothers, 1936.

BIVENS, P. A., *The Ratio Chart in Business*, Norwood, Mass., the Codex Book Co., 1926.

CROXTON, FREDERICK E., and STEIN, HAROLD, "Graphic Comparisons by Bars, Squares, Circles, and Cubes," *Journal of the American Statistical Association*, 27 (177), March, 1932, pp. 54–60.

KARSTEN, K. G., *Charts and Graphs*, New York, Prentice-Hall, 1923.

Management's Handbook, Section 3, "Charts," by D. B. Porter, New York, The Ronald Press Co., 1924.

MUDGETT, BRUCE D., *Statistical Tables and Graphs*, Boston, Houghton Mifflin Co., 1930.

EXERCISES AND PROBLEMS

1. Index numbers of industrial production based on the average of 1929 as 100 indicate the following levels for a number of the principal nations of the world as of April, 1938:

Japan	174.6	Germany	123.5	Poland	92.5
Sweden	146.0	United Kingdom	113.7	France	78.7
Denmark	136.0	Italy	99.9	United States	74.5

Prepare a bar chart that effectively contrasts the given levels of these nations.

2. In 1937, each dollar of revenues received by the Class 1 railways of the United States was expended as follows: for labor, 44.8 cents; for locomotive fuel, 6.3 cents; for other materials and supplies, 17.3 cents; for losses and damages and pensions, insurance, and depreciation, 6.5 cents; for taxes, 7.8 cents; for equipment and joint facility rentals, 3.1 cents; for return to capital, 14.2 cents.

Prepare a circle and sector or pie chart showing this division of operating revenues. Then prepare a composite bar chart for the same purpose.

3. In 1938, the National Association of Manufacturers sought to discover from its members the status of workers over 40 years of age. Some 38.3 per cent of the total returns gave reasons for showing some preference in hiring to workers under 40 years of age. The reasons were divided as follows:

REASONS FOR PREFERRING YOUNGER WORKERS	PER CENT
Training and apprenticeship requirements	31.31
Better work qualifications	30.27
Long-term benefits	22.86
Insurance requirements	8.04
Need for new blood in organization	4.49
Miscellaneous	3.03

Prepare an appropriate graphic representation for these data.

4. Tabulated below is the distribution of railroad operating revenues for a number of years. Prepare a series of charts that effectively portray the changing nature of this distribution.

	1931	1932	1933	1934	1935	1936	1937
Total operating revenues............	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Labor (salaries and wages)*........	46.9	46.0	43.2	44.1	45.0	42.9	44.8
Fuel (locomotive)...............	5.3	5.4	5.1	5.8	5.9	5.9	6.3
Material, supplies, and miscellaneous	17.0	16.2	15.4	15.9	16.2	16.4	
Loss and damage, injuries to persons, insurance, and pensions.........	2.5	2.6	2.5	3.0	2.3	2.3	19.0
Depreciation and retirements......	5.3	6.7	6.5	5.9	5.7	4.8	4.8
Taxes.........................	7.3	8.8	8.1	7.3	6.9	7.9	7.8
Hire of equipment and joint facility net rentals....................	3.2	3.9	3.9	3.9	3.5	3.3	3.1
Total expenses and taxes......	87.5	89.6	84.7	85.9	85.5	83.5	85.8
Net railway operating income........	12.5	10.4	15.3	14.1	14.5	16.5	14.2

* Does not include payroll chargeable to capital account.

5. The following data indicate, in thousands, the number of passenger cars sold to consumers in the United States by the General Motors Corp.:

Year		Year		Year	
1929	1,499	1933	756	1937	1,594
1930	1,058	1934	927	1938	1,002
1931	938	1935	1,279	1939	1,365
1932	510	1936	1,720		

Prepare an ordinary line chart of these figures. Also plot the data on semi-logarithmic paper. What advantages can be claimed for each type of chart?

6. The following data represent the public debt of the United States in each 5-year period since 1850. Chart the figures on semi-logarithmic paper.

Year	Debt (millions of dollars)	Year	Debt (millions of dollars)
1850	63	1895	1,097
1855	36	1900	1,263
1860	65	1905	1,132
1865	2,678	1910	1,147
1870	2,436	1915	1,191
1875	2,156	1920	24,298
1880	2,091	1925	20,516
1885	1,579	1930	16,801
1890	1,122	1935	28,701
		1940, est.	43,000

7. The following tabulation describes the age distribution of the population of the United States according to two censuses:

Age group	Per cent of the total population	
	1900	1930
Under 5 years........	12.1	9.3
5 to 9 inclusive.....	11.7	10.3
10 to 14.............	10.6	9.8
15 to 19.............	9.9	9.4
20 to 24.............	9.7	8.9
25 to 29.............	8.6	8.0
30 to 34.............	7.3	7.4
35 to 39.............	6.5	7.5
40 to 44.............	5.6	6.5
45 to 49.............	4.5	5.7
50 to 54.............	3.9	4.9
55 to 59.............	2.9	3.8
60 to 64.............	2.4	3.1
65 to 69.............	1.7	2.3
70 to 74.............	1.2	1.6
75 to 79.............	0.7	0.9
80 to 84.............	0.3	0.4
85 and over......... ..	0.2	0.2
Unknown............	0.3	0.1

(*a*) Ignoring the last two items, prepare histograms showing these distributions and comparing the proportions in each age group at the two census periods.

(*b*) Prepare a chart of the "less than" cumulative type for each of the two periods.

8. In a study of income in a certain agricultural area, the following distribution was obtained.

Classes (1,000 dollars)	Persons, f, %	Income f, %
0– 2	10	2
2– 4	25	14
4– 6	30	29
6– 8	20	27
8–10	10	17
10–12	5	11

(*a*) Cumulate each frequency expressed as a percentage (*f%*) in a "less than" summation, labeling the first cumulative series *X*, and the second *Y*.

(*b*) Plot on a square cross-section area *Y* against *X*. This is called a Lorenz curve. A diagonal line from 0 on each scale to 100 on each scale is called the "line of equal distribution," and the departure of the plotted curve (*Y* on *X*) from it is a graphic measure of the inequality of income. (See Arkin and Colton, *Graphs: How to Make and Use Them*, p. 65 ff.)

9. Plot as a Lorenz curve the following data of all personal incomes in the United States in 1929, as adapted from *America's Capacity to Consume* (Brookings Institution, Washington, D. C., 1934), p. 207.

Income	Cumulative percentages	
	Persons (*X*)	Income (*Y*)
Under $ 0	0.4	—1.1
500	10.4	0.7
1,000	39.8	12.5
1,500	64.8	28.8
2,000	80.8	43.3
2,500	88.4	52.3
3,000	91.7	57.0
4,000	94.9	62.7
5,000	96.4	66.2
10,000	98.7	74.5
25,000	99 6	82.0
. . . . . .	100.0	100.0

10. The following table (*Monthly Labor Review*, May, 1940, p. 1082) presents indexes of strike data from 1914 to 1939. Draw line charts representing these data.

Year	Index (1927–29 = 100)			Year	Index (1927–29 = 100)		
	Strikes	Workers involved	Man-days idle		Strikes	Workers involved	Man-days idle
1914	162			1927	95	106	178
1915	214			1928	81	101	86
1916	509	514		1929	124	93	36
1917	598	395		1930	86	59	23
1918	451	399		1931	109	110	47
1919	488	1,337		1932	113	104	71
1920	458	470		1933	228	376	115
1921	321	353		1934	250	472	133
1922	149	519		1935	271	359	105
1923	209	243		1936	292	254	94
1924	168	210		1937	637	598	193
1925	175	138		1938	373	221	62
1926	139	106		1939	351	377	121

11. The following data from *The Agricultural Situation* (United States Department of Agriculture) present various indexes of production and prices.

Year and month	Industrial production (1923–25=100)	Income of industrial workers (1924–29=100)	Cost of living (1924–29=100)	(1910–14 = 100)				Farm wages	Prices received by farmers (August 1909–July 1914=100)	Ratio of prices received to prices paid
				Wholesale prices of all commodities	Prices paid by farmers for commodities used in—					
					Living	Production	Living and production			
1925.............	104	98	101	151	164	147	157	176	156	99
1926.............	108	102	102	146	162	146	155	179	145	94
1927.............	106	100	100	139	159	145	153	179	139	91
1928.............	111	100	99	141	160	148	155	179	149	96
1929.............	119	107	99	139	158	147	153	180	146	95
1930.............	96	88	96	126	148	140	145	167	126	87
1931.............	81	67	88	107	126	122	124	130	87	70
1932.............	64	46	79	95	108	107	107	96	65	61
1933.............	76	48	76	96	109	108	109	85	70	64
1934.............	79	61	78	109	122	125	123	95	90	73
1935.............	90	69	80	117	124	126	125	103	108	86
1936.............	105	80	81	118	122	126	124	111	114	92
1937.............	110	94	84	126	128	135	130	126	121	93
1938.............	86	73	82	115	122	124	122	124	95	78
1939.............	105	83	82	113	120	122	121	124	93	77
1939—April.......	92	75	82	111	....	....	120	121	89	74
May.......	92	75	81	111	....	....	120	....	90	75
June.......	98	80	81	110	119	121	120	....	89	74
July.......	101	80	81	110	....	....	120	126	89	74
August.....	103	83	81	109	....	....	119	....	88	74
September..	111	86	82	115	122	123	122	....	98	80
October....	121	91	82	116	....	....	122	126	97	80
November..	124	93	82	116	....	....	122	....	97	80
December..	128	93	82	116	121	123	122	....	96	79
1940—January...	119	93	82	116	....	....	122	119	99	81
February...	109	89	82	115	....	....	122	....	101	83
March.....	103	87	82	114	121	....	122	....	97	80
April.......	102	....	....	115	....	....	123	124	98	80

Prepare charts showing:

(*a*) A comparison of wholesale prices and agricultural living costs.

(*b*) A comparison of income of industrial workers and farm wages.

(*c*) Prices received by farmers, and prices paid for living and production, together with the ratio of the two (ratio chart).

12. Make use of data obtained from business and commercial magazines or reporting services to prepare one of each of the following types of charts:

(*a*) A circle and sector chart.

(*b*) A component bar chart.

(*c*) A multiple-line chart.

(*d*) A statistical map.

CHAPTER V

AVERAGES

When statistical data describing business conditions have been gathered and arranged to permit comparison of their size or magnitude, as described in the preceding chapters, one of the questions that most commonly arises concerns their *average*. Thus, if a study discloses the individual amounts of some four hundred sales in a certain department of a store, a natural question is that of the average size of such sales.

In such cases, the *average* usually connotes the *mean* or *arithmetic mean*, which is sometimes described as a number that is typical of the whole group; that is, it is representative of what is frequently called the *central tendency*. Strictly speaking, it is not always typical of a distribution. If the individual items vary greatly in size, if, for instance, sales include a large number of 10-cent items and a few $100 ones, then the *mean* may be said to express the *summary* character of the whole rather than to typify it.

The arithmetic mean.—The arithmetic mean is calculated as the sum of the items divided by the number of the items. Thus the average of 10, 20, and 24 is 18, which is their total, 54, divided by 3, the number of items. This is the common average, which is familiar to everyone. When the terms "average" or "arithmetic mean" or simply "mean" are used without further qualification in subsequent pages, they refer to this kind of average. Its symbol is either AM or M, and the simplest formula for its calculation is

$$M = \frac{\Sigma m}{N} = \frac{\Sigma X}{N}$$

where m or X (or any other convenient symbol) represents the

measures or items to be averaged, and N is the number of these items.

If the data to be averaged are arranged in classes in a frequency distribution, in effect the same formula is applied. In this case each item in a class is assumed to be represented by the class mark (m). The summation (Σ) of the m's therefore requires that each be multiplied by its frequency, and the formula may be written

$$M = \frac{\Sigma fm}{N} = \frac{\Sigma fX}{N}$$

where the m's or X's represent the measures, or mid-points of the classes, f is the frequency of each measure, and N is the total number of items, or Σf (see Example 5·1, part II). In view of more complex formulas to be used in subsequent operations, it is usually more satisfactory to omit the f, since it is logically implied by the summation sign. Σm indicates a summing of all the m's or X's, and to perform this operation requires that each be included as many times as its frequency indicates. Hence, the abbreviated form, Σm, has the same meaning as Σfm and also has the advantage of applying to cases not involving frequencies.

Deviations from the mean.—It is a characteristic of the arithmetic mean that it is equidistant from the combined items below it and the combined items above it. Each of the individual differences $(d = m - M)$ is called a deviation from the mean when the mean is taken as the origin (R) of these deviations.[1] An item larger than the mean has a positive deviation; one smaller than the mean has a negative deviation. Expressed algebraically, the sum of such deviations is necessarily zero. For example, the deviations (d) of the items 10, 20, and 24, from their average, 18, are:

---

[1] A series of items expressed as deviations from their mean is said to be centered. When each item is expressed by the symbol X, the corresponding deviation (i.e., $X - M_x$) is expressed as x, and $\Sigma x = 0$. However, it should be noted that some statisticians, particularly the English, use x (or other letters) to denote each item in a series, while the mean is written as $\bar{x}$, and a deviation as $x - \bar{x}$.

$$m \text{ or } X \quad M \qquad d \text{ or } x$$

m or X		M		d or x
10	−	18	=	−8
20	−	18	=	+2
24	−	18	=	+6
		Total	=	0

For this reason, the arithmetic mean of a group of items is sometimes defined as the number which, if used to replace each of them, would result in the same sum. It is obvious that this characteristic of the arithmetic mean follows directly from the method of its calculation. That is, the sum of all the deviations, each being $m - M$, is $\Sigma m - NM$, which necessarily equals zero ($\Sigma X - NM = \Sigma x = 0$).

Short-cut calculation of the mean.—Since the sum of the deviations from the arithmetic mean is necessarily zero, a method of computation may be devised which selects a convenient number roughly approximating the mean as an arbitrary starting point, or origin, for the deviations. The deviations (d') from this assumed mean or arbitrary origin (R) are then averaged to obtain a correction figure (c). The actual mean is found by adding the correction figure thus obtained to the arbitrary origin. For example, if the numbers 10, 20, and 24 are to be averaged, 20 may be selected as the arbitrary origin. Then the deviations are -10, 0, and 4. The average of these deviations is $(-10 + 0 + 4) \div 3 = -2$. If this correction is added algebraically to the arbitrary origin, 20, the result, 18, is the corrected or true mean. If the sum of these deviations is zero, then obviously R is the mean. The formula representative of the process thus utilized in discovering the mean may be written as follows:

$$M = R + c = R + \frac{\Sigma d'}{N}$$

where R is the origin that has been assumed, $\Sigma d'$ is the algebraic sum of the deviations from it, and $c = \Sigma d' \div N$.[1] It will be

---

[1] The principle involved in this method of finding the arithmetic mean by assuming an arbitrary origin and making a correction may be explained in another way. The deviations may be considered merely as the original measures, each reduced by R, or in this case, by 20. Hence the average of these reduced measures will be less than the true average by 20, and the true average is $-2 + 20$, or 18.

noted that the symbol d or x has been used to denote a *centered* deviation (one based on the mean), while d' is here employed to distinguish an *uncentered* deviation (one not based on the mean). In later more complex formulas d' will be found inconvenient, and d may be used in its place. But when there is danger of confusing the two types of deviations, either d', D, or some other convenient symbol should be employed for the uncentered deviation.

The same short-cut method may be applied to data that are grouped in the form of a frequency distribution. In this calculation, one of the mid-points (m) is usually selected as an arbitrary origin, and the deviations of the other m's from it, either directly or in units of class intervals, are noted. The deviations thus obtained are multiplied by their frequencies and averaged, and the resulting correction figure, multiplied by the class interval, if this is the unit, is added (algebraically) to the arbitrary origin. The formula for the arithmetic mean of such grouped data may be written

$$ M = R + \frac{\Sigma f d'}{N} \quad \text{or} \quad R + i\left(\frac{\Sigma f d'_i}{N}\right) $$

where d'_i indicates an uncentered deviation expressed in class intervals. The f, however, is written merely as a reminder, and is really implied in the symbol Σ, which means add *all* the items. As was previously suggested, in more complex formulas it is best to omit f entirely, because its use complicates them and sometimes renders them ambiguous.

Calculation of the arithmetic mean in ungrouped data, using both the direct and the short method, is illustrated in Example 5·1, part I. First, the ungrouped data are averaged by the direct method, in which given corn yields are added and divided by the number of experimental plots. The same average is then computed indirectly by assuming an origin, 30, noting the deviations from it, averaging them, and adding this average to the assumed origin.

In Example 5·1, part II, the average of grouped data is similarly computed; and the data and mean are plotted, both

<div align="center">

EXAMPLE 5·1

THE ARITHMETIC MEAN

</div>

I. Ungrouped data; corn yields in bushels per acre, on 5 experimental plots.

A. Direct method	B. By assumed origin, $R = 30$	
m or X	m or X	$d' = m - R$
25	25	-5
30	30	0
33	33	3
37	37	7
45	45	15
$5\overline{)170}$		$5\overline{)20}$
$M = 34$		M of $d' = 4$
		M of $m = 30 + 4 = 34.$

II. Grouped data; tensile strength (test) of 50 sample cords, in pounds, as tested by the purchasing department of the S chain stores.

<div align="center">A. Direct method   B. By assumed origin, $R = 15$</div>

L_1	L_2	m	f	mf	L_1	L_2	m	f	d'	fd'
10	11.99	11	3	33	10	11.99	11	3	-4	-12
12	13.99	13	15	195	12	13.99	13	15	-2	-30
14	15.99	15	20	300	14	15.99	15	20	0	0
16	17.99	17	10	170	16	17.99	17	10	2	20
18	19.99	19	2	38	18	19.99	19	2	4	8
			50	$\overline{)736}$				50	$50\overline{)-14.00}$	
			$M = 14.72$						M of $d' = -0.28 = c$	
									Add R 15.00	
									M of $m = \overline{14.72} = R + c$	

III. Grouped data; deviations expressed in terms of class intervals. Data: Same as section II above.

L_1	L_2	m	f	d_i'	fd_i'
10	11.99	11	3	-2	-6
12	13.99	13	15	-1	-15
14	15.99	R-15	20	0	0
16	17.99	17	10	1	10
18	19.99	19	2	2	4
			50	$50\overline{)-7.00}$	

$$c_i = M \text{ of } d_i' = -0.14 \text{ in terms of class interval}$$
$$i = \underline{2} = \text{class interval}$$
$$c = -0.28 = \text{correction}$$
$$M = R \text{ plus correction} = 15 - 0.28 = 14.72.$$

as a frequency distribution and a cumulative curve, in Figs. 5·1 and 5·2. In the direct method, IIA, each mid-class or class mark is multiplied by its frequency. The total thus obtained is then divided by the sum of the frequencies, or N. The same result is obtained in part IIB, by assuming an origin approximating the mean and near the center of the distribution where the frequencies are large, after which the deviations from this origin are averaged, and their average, -0.28, is added to the assumed origin.
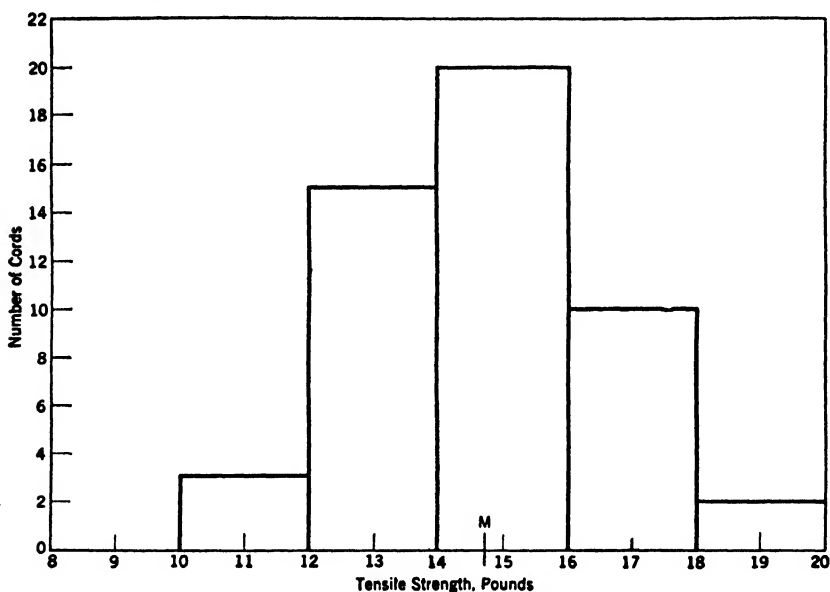
FIG. 5·1.—Rectangular Chart or Histogram. Data: Example 5·1, part II.

In problems involving large numbers of items spread over a wide range, and particularly where fractional values are involved, it is more convenient to make use of an origin at some convenient m, and class unit deviations from it. This method is illustrated in section III of the example. It differs from part IIB, only in that deviations are expressed in terms of class intervals (d_i') as -1, -2, etc., and 1, 2, etc., instead of in terms of the actual deviations. In the calculation of the mean, it is necessary to multiply the average of the deviations $(\Sigma f d_i' \div N)$

by the amount of the class interval (i), before adding this correction figure to the assumed origin. Obviously this method is convenient only when the class interval is constant throughout the distribution.

Frequencies and weights.—An average of tabulated data such as that calculated in Example 5·1, parts II and III, is sometimes called a weighted average. That is, it is the average of the m's, with greater or less emphasis given to each of them
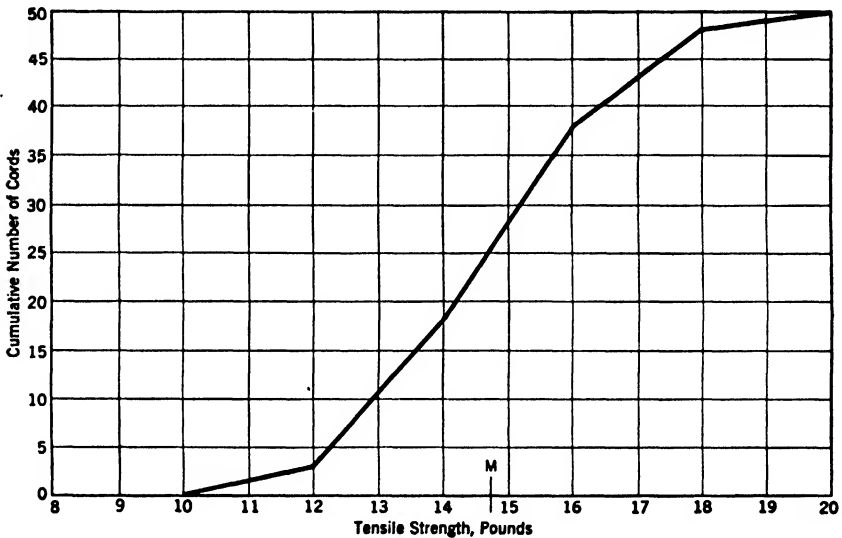
FIG. 5·2.—Cumulative Frequencies of Example 5·1, part II.

according to its frequency. From this point of view, the frequencies appear as weights, and the formula becomes:

$$M = \frac{\Sigma mw}{N} = \frac{\Sigma Xw}{N}$$

where w signifies a weight, or, in effect, a frequency. Sometimes, although not often, frequencies and weights are required in the same computation of a mean. When this occurs, the product of the frequencies and weights for each m is applied as a weight to that m.

The term weight, however, is somewhat broader than the

term frequency. There are cases where the former term is more appropriate, although it has practically the same significance as a frequency. For example, suppose that two consignments of grain were received, one of 1,000 bushels and the other of 3,000 bushels, at prices of $0.90 and $1.10 per bushel, respectively. The average price would be found thus:

PRICE	QUANTITY	PRODUCT
$0.90	1000 bu.	900
1.10	3000 bu.	3300
	4000 bu.	4000)4200(1.05

The average, then, is $1.05 per bushel. The quantities, 1,000 and 3,000, might be called *frequencies*, since they indicate how many times the price is spent, but the term *weights* is preferable.

It will be seen from the preceding example that the weights might be taken as 1 and 3 instead of 1,000 and 3,000, without changing the result. *Weights or frequencies may be multiplied or divided through by a constant without affecting the resulting average*, for it is the ratio of weights or frequencies to each other and not their absolute magnitudes that is significant.

It should be noted that the arithmetic mean will sometimes ⸙ be inappropriate when the data are not homogeneous, that is, when they represent a mixture of two or more differing classifications. For example, the average wage of a group of unskilled workers, combined with another group of highly paid experts, would not be meaningful, though the mean of the graded incomes in a typical community might be significant. Methods for determining whether subgroups differ significantly among themselves will be discussed later.[1]

The geometric mean (GM, G, or M_g).—Somewhat analogous to the arithmetic mean as a measure of central tendency is the *geometric mean*. The geometric mean of N numbers is the Nth root of their product. It may also be calculated in a manner similar to that by which the arithmetic mean is discovered except that the measures are transferred to the geo-

---

[1] See Chapters VII and XIX.

metric scale by using their logarithms instead of the measures themselves.   Tables of logarithms, such as are included in the Appendix, may be used for this purpose, or logs and antilogs may be read from a slide rule.   In many cases, the actual multiplication and division may be conveniently performed on the slide rule without direct resort to logarithms.

In general, it may be said that the geometric mean is called for when the items are considered as factors or ratios.   As an example of the method of calculation, the geometric mean of 4 and 9 is found as the square root of the product of the numbers, or, $GM = \sqrt{4 \times 9} = 6$.   Or, employing logarithms, as must be done in more complex problems,

$$
\begin{array}{r}
\log 4 = .602 \\
\log 9 = .954 \\
\hline
2\overline{)1.556} \\
\hline
.778
\end{array}
$$

antilog $.778 = 6$

The geometric mean of 9 and 4, therefore, is 6.   The significance of this figure may be illustrated by reference to the fact that a surface 4 feet by 9 feet may be said to have an average dimension of 6 feet because $4 \times 9 = 6 \times 6$.   That is, 6 is the number used to replace 4 and 9 without changing the result under the conditions of the problem.

Another situation in which the geometric mean is indicated arises when an average rate of successive increases and decreases is required.   Suppose, for example, the population of a certain city grew 50 per cent in one decade, 25 per cent in the next decade, and declined 5 per cent in the third decade, and the average rate of increase per decade is required.   If P stands for the initial population, then the population at the end of the period is indicated as $P \times 1.50 \times 1.25 \times 0.95 = 1.78125P$.   In this calculation 1 plus each rate $(1 + r)$, taking the rate with its algebraic sign, are factors, respectively, in the final product. The geometric mean of $1 + r$ is $\sqrt[3]{1.50 \times 1.25 \times 0.95} = \sqrt[3]{1.78125} = 1.212$.   That is, the average rate of increase is

0.212, or 21.2 per cent. This result may be checked by substituting it in the original formula thus:

$$P \times 1.212 \times 1.212 \times 1.212 = 1.78P$$

Since the geometric mean of N numbers is the Nth root of their product, the geometric mean of three numbers may be found as the cube root of their product.[1] For example, a room having

EXAMPLE 5·2

THE GEOMETRIC MEAN

Data: See Example 5·1, page 91.

I. Ungrouped data

m or X	$\log m$ or $\log X$
25	1.3979
30	1.4771
33	1.5185
37	1.5682
45	1.6532

$$5)\overline{7.6149}$$
$$\log GM = 1.5230$$
$$GM = 33.3$$
$$\log GM = \Sigma(\log m) \div N = 7.6149 \div 5 = 1.5230$$
$$GM = \text{antilog } 1.5230 = 33.3.$$

II. Grouped data

$L_1 - L_2$	X	$\log X$	f	$f \times \log X$
10−12	11	1.0414	3	3.1242
12−14	13	1.1139	·15	16.7085
14−16	15	1.1761	20	23.5220
16−18	17	1.2304	10	12.3040
18−20	19	1.2788	2	2.5576

$$50 \qquad 50)\overline{58.2163}$$
$$\log GM = 1.1643$$
$$GM = 14.60$$

[1] For a description of convenient procedure in machine calculation of square and cube roots, see the instructions published by the Marchand Calculating Machine Co., Oakland, Calif.

the dimensions $10 \times 25 \times 32$ would have an average dimension of $\sqrt[3]{10 \times 25 \times 32} = 20$.

It is obviously advantageous to use logarithms in computing the geometric mean except in simple cases where tables of squares and cubes can be used. Of course, it is possible to find the fourth root by taking the square root of the square root, or the sixth root as the cube root of the square root. But, in general, the use of logarithms is preferable in the more complex cases.

The calculation of the geometric mean by logarithms is illustrated for both ungrouped and grouped data in Example 5·2. The numbers to be averaged, that is, the m's or X's, are replaced by their logarithms.[1] The arithmetic mean of these logarithms is then found. This mean is the logarithm of the geometric mean, which may be found as its antilogarithm in a logarithmic table. The short-cut method using an assumed mean is not⋅ convenient in this case because the deviations from the assumed mean, when stated in logarithms, are as difficult to average as the original logarithms themselves.

The harmonic mean (HM).—The *harmonic mean* is not a widely used form of average, but it must be given some consideration, because it is appropriate in certain types of problems involving weighted averages. The nature of the harmonic mean and what is perhaps the simplest way in which to approach it may be illustrated by the following data, which may be assumed to represent essential facts regarding two purchases of a certain commodity, and in which the problem is the discovery of the average price in the two transactions:

DATE	PRICE PER POUND	TOTAL COST
June 1, 1939	$0.10	$20.00
June 2, 1939	0.15	45.00

[1] To obtain log of 25, find log of 2.5 in table (.3979), and prefix 1 because 25 is in "tens." Similarly obtain other logs, and calculate weighted average, which is the log of GM. To find antilog, locate number nearest to .5230 in body of table (.5224), and read 3.3 in margin, and annex the 3 at the head of column. Read GM as 10 times 3.33 because log GM has a 1 preceding decimal, indicating "tens." With a larger log table, the result may be read to more decimal places. It may be noted that if a distribution is of the logarithmic type (normal on a logarithmic X scale) the geometric mean is equivalent to the median, to be described later.

It might appear, at first glance, that prices featuring each purchase should be weighted by the total cost of each purchase. But price multiplied by total cost does not result in a reasonable or sensible product. Rather, the natural weight for the price per pound is the number of pounds, since the number of pounds represents the number of times the price is spent and is, therefore, the frequency. For this reason, in order to secure an appropriate weighted average, it is necessary to determine the number of pounds purchased on each of the dates. This result is achieved by dividing $20.00 by $0.10 and $45.00 by $0.15, thus discovering 200 and 300 pounds, respectively, as the appropriate weights. The problem may then be restated, with the purchases described as follows:

DATE	PRICE PER POUND	POUNDS	TOTAL COST
June 1, 1939	$0.10	200	$20.00
June 2, 1939	0.15	300	45.00
		500	$65.00

Such a restatement makes clear the average price per pound as the total cost, $65.00, divided by the total number of pounds, 500, or $0.13 per pound.

Problems of the type just described, in which rates must be averaged, involve calculation of what is known as the *harmonic mean*. That measure may be described as the reciprocal of the arithmetic mean of reciprocals of the measures. Frequencies in such a calculation are of the type illustrated by "total costs" in the illustration above. As is implied in the definition of the harmonic mean, a common procedure in its calculation involves the discovery of the reciprocals of the class measures, the average of these reciprocals, and the reciprocal of this average. The last-mentioned measure is the harmonic mean. The process may be illustrated, utilizing the data described above, as follows:[1]

[1] It will be clear that calculations may be reduced by omitting the columns of reciprocals and obtaining the product $(1/X)f$ directly as f/X, and, similarly, by calculating HM directly as $\dfrac{N}{\Sigma(f/X)}$, without reference to the M of reciprocals. See, for a discussion of the harmonic mean, E. F. Ferger, "The Nature and Use of the Harmonic Mean," *Journal of the American Statistical Association*, 24 (173), March, 1931, p. 36.

Purchase, or Date	Price per Pound	Reciprocal	Frequency	Product
	X	$1 \div X$	f	$(1 \div X)f$
I. June 1, 1936	$0.10	10	20	200
II. June 2, 1936	0.15	6.667	45	300
			65	500

$$M \text{ (of reciprocals)} = 500 \div 65 = 7.692$$
$$HM = 1 \div 7.692 = 0.13$$
$$\text{Or, } HM = 65 \div 500 = 0.13$$

AVERAGES OF POSITION

Arithmetic, geometric, and harmonic means described in earlier paragraphs of this chapter are strictly mathematical measures, in which each item in a distribution plays a part determined by its listed magnitude. There are, however, several measures of central tendency that are frequently useful, although they do not have a similar mathematical relationship to the data of the distributions they represent. Attention may now be directed to two such averages that depend on the position of items in an array or frequency distribution. The term "array" means that items have been arranged *in the order of their size,* from the smallest to the largest, or, sometimes, in the reverse order.

The usefulness of such measures to business data may be somewhat more apparent if mention is made of a few typical uses. Assume, for instance, that the investment division of a business organization has collected data showing the rates of interest paid by various securities held by the corporation, and that it is desired to designate the most representative of the lot, so that certain of them could be isolated for further investigation. In such a case, issues would probably be arrayed according to their earning power, and the central item in the array, the *median,* selected as the most representative, while issues above or below such a position average might be studied further.

Similarly, various types of fuel might be arrayed according to the cost per horsepower derived from each of them. If their average or mean did not actually represent any individual item, their median might reasonably be preferred to any statement of their average as a representative measure of the group.

In another aspect of business, that which compares measures of production or sales or other business activity from month to month over a period of years, it is frequently found that seasonal data representing a given month in several years, when brought together and arrayed, display a pattern in which most items are fairly close together while a few of them, for a variety of reasons, such as unusual climatic or industrial conditions, are irregular and unusual. In such cases, the median item is likely to be more representative of these data than their mean, particularly from the standpoint of its usefulness in forecasting. When the data are numerous, the mode or most common magnitude may prove a more satisfactory measure. In still other cases, the location of quartiles, deciles, or percentiles (all of which are described in later paragraphs of this chapter) may be most useful.

The median (*Md*).—The median may be defined as the item, actual or interpolated, occurring at the central position in an array. For example, if a group of 5 investments earn rates of 3, 3.5, 3.75, 4.25, and 6 per cent respectively, the median return is the item occupying the central position in the array, namely 3.75 per cent. It will be noted that the median thus determined does not take account of the actual size of the items other than the one selected as the median, account being taken only of their position in the array. The arithmetic mean, on the other hand, is affected by every measure in the series. In this case, the mean is 4.1, which happens to be considerably larger than the median. If there had been an even number of items in the array, the median would have been taken as the average of the two central items. Thus, in the array: 3, 3.5, 3.75, 4.25, 6, and 8, the median is the average of 3.75 and 4.25, which is 4.

When a long list of items is arrayed, the middle position may also be readily determined by counting successive spaces between items to the space number $N \div 2$. For example, in 100 items so arrayed, the median position is item number $50\frac{1}{2}$ or space number 50, both of which indicate the average of the fiftieth and fifty-first items. The latter is probably more convenient as a method, since it is applicable to frequency distri-

butions, where class limits fall, at least theoretically, between items.

If the median is to be estimated on the basis of a frequency tabulation without recourse to the original numbers from which the tabulation is made, it must be assumed that the given numbers are representative of an infinitely large smoothed distribution. The process of so-called linear interpolation commonly employed in locating the median under such circumstances is comparatively simple, and may be described as follows:

If the data are written to show the class limits and the corresponding summations of the frequencies at the beginning and ending of each class, the first step in discovering the median is the location of the class in which it falls. Then it is necessary to consider where, in this class, the median item is found. Thus if 20 items are classified as follows (L as usual refers to class limits):

CLASS	FREQUENCIES	CUMULATIVES
L_1–L_2	f	Σ_1–Σ_2
16–20	6	0– 6
20–24	8	6–14
24–28	6	14–20
	$N = 20$	

then it is apparent that (1) the tenth space is the median space since $N \div 2 = 10$, (2) the second class is the median class since 10 falls between the summation at the beginning of the class and the summation at the end of the class, and (3) the median space is half of the way through this class since 10 is half of the way from 6 to 14, a distance represented by the class frequency; that is, $(10 - 6) \div 8 = \frac{1}{2}$. Hence the median is this fraction ($\frac{1}{2}$) times the class interval, (4), plus the lower limit of the class (20), or $(\frac{1}{2} \times 4) + 20 = 22$. In other words, since ($N \div 2$) is half way from $\Sigma_1 = 6$ to $\Sigma_2 = 14$, the median is half way from $L_1 = 20$ to $L_2 = 24$. Or, in general

$$Md = \left(\frac{\dfrac{N}{2} - \Sigma_1}{f} \times i \right) + L_1$$

applied to the class in which the median is located.   The computation is illustrated in detail for both ungrouped and grouped data in Example 5·3.

<div align="center">EXAMPLE 5·3</div>

<div align="center">THE MEDIAN</div>

Data: See Example 5·1, page 91.

I. Ungrouped data

   Data "arrayed" by size: 25   30   33   37   45   55
   Spaces between items, number: 1   2   3   4   5
   Md position at space $N \div 2 = 3$, or 3d space

$$Md = (33 + 37) \div 2 = 35$$

II. Grouped data

Class limits L_1–L_2	Frequency f	Cumulatives Σ_1–Σ_2
10–12	3	0– 3
12–14	15	3–18
14–16	20	18–38 (Md class)
16–18	10	38–48
18–20	2	48–50
$i = 2$	50 $N/2 = 25$	

$$Md = \left(\frac{\frac{N}{2} - \Sigma_1}{f} \times i \right) + L_1$$

$$= \left(\frac{25 - 18}{20} \times 2 \right) + 14 = 14.70$$

The median as a measure of central tendency is well adapted to distributions where, for one reason or another, it seems desirable to give only slight emphasis to extreme items.   Such cases occur very frequently in statistical procedure, but the choice of the median is largely a matter of judgment rather than of exact mathematical procedure.   Thus, in items representing measurements, occasional extreme items may be regarded as probably erratic or abnormal, and the median may be used to discount

their influence. For example, suppose that the average or typical weekly performance of a given machine is required, and that its output was greatly restricted by accidental circumstances for one or two weeks. In such a case, the median performance may be more typical than the mean.

Sometimes a compromise is made between a mean and a median. For example, the series, 85, 92, 95, 98, 99, 100, 102, 103, 110 might be represented by the average of the three median items, 98, 99, and 100. This procedure involves striking out the three largest items and the three smallest items. When this is done the number of items struck out at each end is the same, and the numbers left represent about a third of the array. This method will be encountered later in computing seasonal variations. The purpose is to give the median a broader base, yet to discount the extremes. Sometimes, again, weights arbitrarily stressing the central items are employed. Such devices may be useful, but they cannot be given a strictly mathematical basis.

Quartiles, deciles, and percentiles.—The *quartiles* are the three points in the range of a frequency distribution that divide it into four parts, each of which contains one-fourth of the total number of items in the distribution. The *deciles* are the nine points that similarly divide the distribution into ten equal classes of items, and the *percentiles* represent the points that divide the distribution into one hundred equal parts. The tenth percentile is thus the first decile, and the twenty-fifth percentile is the first quartile. In the same way, the fiftieth percentile is the fifth decile and the median of the distribution. The range between two quartiles is described as the *quartile interval*. Methods of discovering the measures of quartiles and percentiles are discussed in the next succeeding chapter.

The mode (*Mo*).—Another commonly used measure of central tendency, applicable to fairly regular frequency distributions, is called the *mode*. As the term suggests, the mode is simply the most common magnitude, actual or interpolated, in the distribution.[1]

---

[1] In a perfectly symmetrical distribution, of course, the mean, median, and mode coincide. In slightly skewed distributions, particularly of the type described as logarithmic normal distributions, it may be shown that the relationship of the mean,

Under some circumstances, the mode may be more useful as a measure of central tendency than the mean or median. If, for example, a sales campaign is being planned which is intended to reach the population of a given city or other area, then the modal income of the group would serve as the best guide to the class of goods to be advertised. Similarly, the modal quality of goods, the modal price in a market, or the modal efficiency of a group of employees may frequently be more useful than any other measure of central tendency, because the mode may accurately typify the whole array. Like the median, the mode is readily calculated for open-class distributions, a characteristic that adds distinctly to its usefulness.

The mode is approximated as the mid-point of the class having the greatest frequency. In a price distribution, for instance, the mode may be taken as the most common price class. It may be readily located on a chart of a frequency distribution as the point on the X scale representing the mid-point of the highest frequency rectangle, which pictures the *modal class*.

It will be clear that, if data were sufficiently numerous and class intervals were made smaller and smaller, the mode could be quite accurately located by charting. In most distributions, however, data are insufficient and classes too limited to permit accurate location of the mode by observational methods. Hence, an attempt is sometimes made to estimate the position of the mode by interpolation within the modal class. Such interpolation is most commonly based upon the relationship between the modal class and the frequencies of adjacent classes. Thus, it will be found that, if a chart is constructed representing the frequencies of the various classes as rectangles, and if a smooth curve is drawn through the tops of rectangles representing the

median, and mode, expressed in terms of their logarithms, is as follows:

$$M - Md = \tfrac{1}{2}(Md - Mo)$$
$$M = \tfrac{1}{2}(3Md - Mo); \; Md = \tfrac{1}{3}(2M + Mo)$$
$$Mo = 3Md - 2M$$

These relationships prevail whether the distribution is positively or negatively skewed, and they are sometimes applied as approximations to a variety of slightly skewed distributions, utilizing the actual rather than the logarithmic measures.

modal class and the two adjacent classes directly above the respective class marks, the highest point of the curve may be regarded as a fair estimate of the position of the mode. This position may be approximated without actually drawing the curve by means of the formula [1]

$$Mo = \left(\frac{d_1}{d_1 + d_2}\right) i + L_1$$

where d_1 and d_2 are the differences between the modal frequency and the preceding and following frequencies, respectively; i is the class interval; and L_1 is the lower limit of the modal class. If, for instance, the three frequencies involved are 15, 20, and 10, respectively, in which 20 is the modal frequency,

$$d_1 = 20 - 15 = 5$$

$$d_2 = 20 - 10 = 10$$

and

$$\frac{d_1}{d_1 + d_2} = \frac{5}{5 + 10} = \frac{1}{3}$$

$$Mo = (\tfrac{1}{3}) i + L_1$$

A little experimentation will show that this method locates the mode nearer to the larger of the two adjacent frequencies, as would be expected, or in this case at a point one-third of the way through the modal class. This process of interpolation is illustrated in Example 5·4. Further, it may be added that, if there are two equal modal frequencies, the class limit common to both is generally taken as the mode.

It should be emphasized that the mode has most meaning when the distribution is regular, i.e., has only one peak fairly

---

[1] Sometimes it is suggested that the mode may be located by the formula:

$$Mo = \left(\frac{f_2}{f_1 + f_2}\right) i + L_1$$

where f_1 and f_2 are the frequencies preceding and following the modal class, respectively, classes being arrayed in the order of their magnitude. This formula has no adequate mathematical foundation, however, and it is lacking in flexibility, as will be apparent if one of the adjacent frequencies is nearly equal to that of the modal class. The one based on differences, however, locates the mode of a regular curve (parabola) passing through the three central frequencies above the m of each.

near the center of its range.   Sometimes, however, distribu-
tions are said to be *bimodal* when they show two peaks.   In any

EXAMPLE 5·4

INTERPOLATING THE MODE

Data: See Example 5·1, Part II, page 91.

Class limits L_1-L_2	Frequency f
10–12	3
12–14	15
14–16	20 (modal class)
16–18	10
18–20	2

Modal class is 14–16

$$d_1 = 20 - 15 = 5.$$
$$d_2 = 20 - 10 = 10.$$

$$Mo = \left(\frac{d_1}{d_1 + d_2} \times i\right) + L_1 = \left(\frac{5}{5 + 10} \times 2\right) + 14 = 0.67 + 14 = 14.67$$

case, the method of interpolation described above provides only
a rough approximation of the mode.   Strictly speaking, there
is no determinable mode in a frequency distribution.[1]

Graphic determination of median and mode.—Both the
median and the mode as interpolated in grouped data may

---

[1] A method based on curve fitting, adapted to distributions with class intervals of
equal size throughout the distribution, may be briefly described.   In principle, it
consists of fitting the skewed binomial $(p + q)^n$, as explained in Chapter XX, and
approximating the mode of this fitted binomial. · The formula is

$$Mo = m_0 + \frac{1}{2n}[2(M - m_0)(n + 1) - in]$$

where m_0 is the first (smallest) m, n is 1 less than the number of classes, and the other
symbols are as usual.   As applied to the following data,

$$m = 2,\ 4,\ 6,\ 8,\ 10,\ 12,\ 14;\quad n = 7\text{-}1 = 6$$
$$f = 3,\ 9,\ 10,\ 8,\ 6,\ 3,\ 1;\quad N = 40$$

where $M = 6.9$ and $n = 6$, the calculation is

$$Mo = 2 + \frac{1}{2 \times 6}[2(6.9 - 2)(6 + 1) - (2 \times 6)] = 6.72$$

be estimated from an appropriate chart, often with sufficient accuracy for practical purposes. Estimating the median is most readily accomplished by means of a cumulative curve, or ogive (see Fig. 6·1, p. 124), while estimation of the mode is facilitated by use of a rectangular frequency chart (see Fig. 5·3). The methods employed will be clear from the construction of the figures. The median is simply the point on the magnitude scale
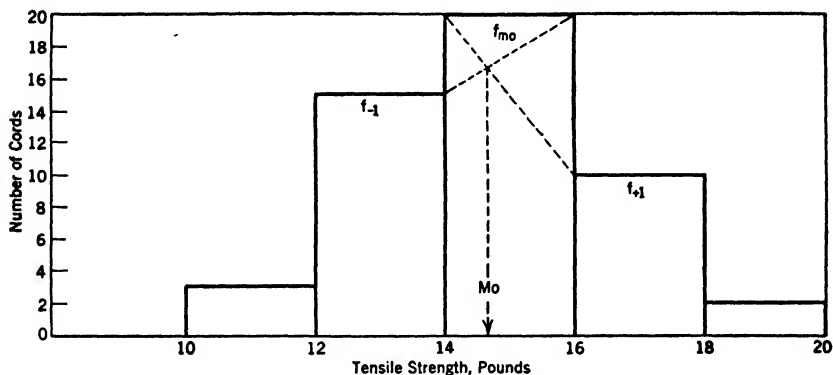
FIG. 5·3.—Graphic Estimation of the Mode.

(X) coordinate with the mid-point of the cumulative scale (Y), as determined by the ogive. The mode, as located on a histogram, is the ordinate where diagonals connecting upper corners of the modal class with the nearest corners of adjacent classes intersect.

READINGS

EXERCISES AND PROBLEMS

A. EXERCISES

1. In the given distributions find the following measures of central tendency: M, Md, and Mo.

(a)		(b)		(c)		(d)		(e)	
m	f	m	f	m	f	m	f	m	f
2	3	4	2	1	1	2	1	2	1
3	5	6	4	2	3	4	2	4	4
4	6	8	6	3	5	6	5	6	6
5	4	10	5	4	2	8	3	8	4
6	2	12	3	5	1	10	1	10	1

2. Calculate the mean, median, and mode, of each of the following distributions:

(a)		(b)		(c)		(d)		(e)		(f)		(g)		(h)		(i)	
m	f	m	f	m	f	m	f	m	f	m	f	m	f	m	f	m	f
5	1	6	2	3	20	2	10	8	30	4	4	10	3	4	4	6	2
15	4	18	4	5	50	4	40	10	60	8	7	20	7	8	6	10	5
25	3	30	3	7	40	6	50	12	50	12	5	30	6	12	5	14	6
35	2	42	1	9	10	8	20	14	20	16	3	40	3	16	3	18	4
										20	1	50	1	20	2	22	3

3. Calculate the mean, median, and mode, of each of the following distributions:

(a)		(b)		(c)		(d)		(e)		(f)	
m	f	m	f	m	f	m	f	m	f	m	f
2	3	3	1	3	1	2	2	3	1	6	2
4	9	5	6	5	7	4	5	4	14	8	12
6	10	7	7	7	11	6	7	5	25	10	24
8	8	9	7	9	9	8	6	6	27	12	25
10	6	11	4	11	6	10	5	7	18	14	17
12	3	13	3	13	4	12	3	8	9	16	10
14	1	15	2	15	2	14	2	9	4	18	7
								10	2	20	3

4. Find the mean, median, and mode of each of the following distributions; also the geometric and harmonic means:

(a)		(b)		(c)		(d)		(e)	
X	f	X	f	X	f	X	f	X	f
20	1	20	1	2	4	1	1	10	2
40	5	40	4	4	7	2	4	12	5
60	3	60	6	6	5	3	5	14	6
80	1	80	4	8	3	4	3	16	4
		100	1	10	1	5	2	18	2
						6	1	20	1

B. PROBLEMS

5. Calculate appropriate measures of central tendency (M, Md, Mo), for the exercises listed at the end of the next chapter, pages 138–146.

CHAPTER VI

DISPERSION

When a series of items, either grouped or ungrouped, has been analyzed to secure some measure of its average, it is frequently desirable to secure one or more measures of the *scatter* of the items, in part in order to determine the degree to which the average is representative of the whole series. The average is clearly more representative if all the items are close to it. The degree to which they scatter from the average or other measure of central tendency is called the *dispersion* or *variation* of the series. It may be measured in several ways, the most important of which require attention at this point.

The range.—The crudest measure of dispersion is the *range.* It represents the distance from the smallest value to the largest value in a given sample. Data may be arranged in an *array,* i.e., they may be recorded in the order of their size or magnitude from smallest to largest. The range is then readily apparent as the difference between the magnitude of the smallest and largest items, or of smallest and largest class limits in grouped data with specific items unavailable.

The average deviation $(A D)$.—Aside from the *range* of items, the simplest measure of dispersion is the *average deviation,* which is the arithmetic mean of differences between individual items and the measure of central tendency, usually the mean. Sometimes the average deviation is measured from the median. The average deviation is also sometimes called the *mean deviation.* When measured from the median, it is theoretically a minimum. To illustrate the nature of the average deviation, reference may be made to a series of three prices, \$10, \$20, and \$24, the common average of which is \$18. The deviations of the items from this average are respectively -8, $+2$, and $+6$. If these devia-

tions are regarded as absolute (i.e., if their algebraic signs are disregarded), their mean is their sum, 16, divided by their

<div style="text-align:center">

EXAMPLE 6·1

THE AVERAGE DEVIATION

</div>

Data: See Example 5·1, page 91.

I. Ungrouped data

Data m or X	$m - M$ d or x
25	-9
30	-4
33	-1
37	3
45	11
5)170	5)28 $= \Sigma|d|$
$M = 34$	$AD = 5.6$

$$AD = \Sigma|d| \div N = 28 \div 5 = 5.6$$
$$\text{Coef. } AD = 5.6 \div 34 = 0.16, \text{ or 16 per cent}$$

II. Grouped data [1]

L_1–L_2	m or X	f	mf	d or x	fd
10–12	11	3	33	-3.72	-11.16
12–14	13	15	195	-1.72	-25.80
14–16	15	20	300	0.28	5.60
16–18	17	10	170	2.28	22.80
18–20	19	2	38	4.28	8.56
		$N = 50$	50)736		-36.96
			$M = 14.72$		$+36.96$
					50) 73.92 $= \Sigma f|d|$
					$AD = 1.4784$

$$AD = \Sigma f|d| \div N = 73.92 \div 50 = 1.48$$
$$\text{Coef. } AD = 1.4784 \div 14.72 = 0.10, \text{ or 10 per cent}$$

---

[1] The average deviation is sometimes calculated with the median as origin. For the above data, if $R = Md$, $AD = 1.484$ (usually minimum AD).

number, 3. That is, $AD = 16 \div 3 = 5.33$ (dollars). This average deviation is a measure of the dispersion or "scatter" of the items. If all of them had been close to the mean, the average deviation would obviously have been smaller. For example, the numbers 17, 18, and 19 have the same mean, i.e., 18, but their average deviation is only 0.67. Since the "scatter" is much less, the average deviation is, of course, also smaller, and the mean is more nearly representative of the group.

In grouped data, the average deviation is obtained by (1) finding the mean, (2) noting the absolute deviation of each class mark from this mean, and (3) averaging these deviations (always taking account of the frequencies involved). The process of calculation as applied to both grouped and ungrouped data is illustrated in Example 6·1. The symbol for d absolute is $'d'$ or $|d|$.

In Example 6·1, part II, it will be seen that the average deviation is 1.48 as measured from a mean of 14.72. This is a way of stating that the distribution spreads through the range, 14.72 ± 1.48, or from 13.24 to 16.20. This range does not include the whole distribution, which ranges from 10 to 20, but it measures the spread by combining the larger variations with the smaller variations. As a rule, it would be necessary to measure three or four average deviations below the mean and about the same amount above the mean in order to include the whole distribution. Thus $14.72 \pm 3 \times 1.48$ represents a spread from 10.28 to 19.16 which includes almost all the items, while $14.72 \pm 4 \times 1.48$ represents a spread from 8.80 to 20.64, which extends beyond the lower and upper limits (10 and 20) of the distribution.

Short-cut *AD*.—Probably the most convenient short cut in calculating the average deviation is one which, when the mean is the origin, relies on the fact that the negative and positive deviations are equal. Thus in Example 6·1, part I, the negative deviations, listed as absolute, may be written

$$
\begin{aligned}
|d_1| &= 34 - 25 = 9 \\
|d_2| &= 34 - 30 = 4 \\
|d_3| &= 34 - 33 = 1 \\
\Sigma &= \overline{102} - \overline{88} = \overline{14}
\end{aligned}
$$

and the total deviations are twice 14, or 28. This sum is divided by N, to secure the average deviation. The process may be condensed by taking the mean times N_s (where N_s is the number of items *smaller* than the mean) less the sum of these smaller items $(25 + 30 + 33 = 88)$, that is,

$$AD = 2(N_s M - \Sigma m_s) \div N$$

$$= 2[(3 \times 34) - 88] \div 5 = 2(102 - 88) \div 5 = 5.6$$

The formula is readily applied to grouped data (part II) by noting that N_s and Σm_s imply the frequencies, that is,

$$AD = 2(N_s M - \Sigma m_s) \div N$$

$$= 2[(18 \times 14.72) - 228] \div 50 = 1.4784$$

where N_s is the sum of the frequencies 3 and 15, and Σm_s is the sum of the mf's 33 and 195. By marking the position of the mean in the m column, these sums are readily obtained.

It should be noted that, if another origin (R) is to be used instead of the mean as the base of the deviations, the formula becomes

$$AD = \frac{2(N_s R - \Sigma m_s)}{N} + M - R$$

where the subscript s refers to R. That is, N_s means the number of items smaller than R, and Σm_s is their sum. Thus in Example 6·1, part II, if the median, \$14.70, is taken as the origin,

$$AD = \frac{2[(18 \times 14.70) - 228]}{50} + 14.72 - 14.70 = 1.484$$

which may be easily verified by direct calculation.

Coefficient of average deviation.—When the dispersion of items in different kinds of measurements is to be compared, a coefficient expressing the *relative* scatter becomes necessary. For example, suppose that the dispersion of the wages of a group of workers is to be compared with the dispersion of scores made by the same workers in mental tests. In such a case, the *coefficient of average deviation* for each dispersion may be used to facilitate comparison. This coefficient is the average deviation

just described expressed as a fraction of the arithmetic mean; i.e., it is the AD divided by the measure of central tendency. For the three items previously mentioned, 10, 20, and 24, the coefficient of average deviation would be

$$\text{Coef. } AD = \frac{AD}{M} = \frac{5.33}{18} = 0.30, \text{ or a 30 per cent scatter}$$

The method of calculating this coefficient is further illustrated in Example 6·1.

The standard deviation (SD or σ).—As a measure of the scatter of grouped or ungrouped items, the average deviation is often used as an informal measure, but it has certain mathematical disadvantages. In the first place, the summing of the deviations without regard to their algebraic signs is an obstacle to more complex mathematical formulas involving dispersion. Hence, if additional work is to be done, another measure of dispersion must be devised. Mathematicians are agreed that the most suitable of such measures is the so-called *standard deviation* (symbol SD, σ, or s).

The standard deviation may be calculated most simply by squaring the deviations of each of the items from their mean (the standard deviation is always measured from the mean), averaging the squares, and taking the square root of this average.[1] The result thus obtained is sometimes called the *root mean square of the deviations*—a term which suggests the three principal operations. It is also called the *quadratic mean of the deviations*—the term quadratic implying a square. In applying this measure of dispersion to the three items, 10, 20, and 24, the mean of which is 18, we first square each of the deviations, -8, 2, and 6. The squares of these deviations are 64, 4, and 36,

---

[1] It is hardly possible at this stage of the discussion to explain why the mathematician prefers the standard deviation to the average deviation as the measure of dispersion. One advantage has already been suggested, namely, that the process of squaring eliminates the minus signs of deviations and thus makes the standard deviation more satisfactory in complex mathematical computations. Further, the final step, the taking of the square root, gives an answer which may be regarded as either plus or minus, that is, as measuring the deviations both above and below the mean. In more advanced analysis, several additional advantages of the standard deviation become obvious.

respectively, the average of which is 34.67. The square root of 34.67 is approximately 5.89, which is the standard deviation. In working with grouped data, the same rule is carried out by finding the deviations of the class marks from the mean, squaring and averaging these deviations ($\Sigma f d^2 \div N$ or $\Sigma x^2 \div N$) and taking the square root. The procedure is applied to both ungrouped and grouped data in Example 6·2.

EXAMPLE 6·2

THE STANDARD DEVIATION—DIRECT METHOD

Data: See Example 5·1, page 91.

I. Ungrouped data

Data m or X	$X - M = d$ or x	d^2 or x^2
25	−9	81
30	−4	16
33	−1	1
37	3	9
45	11	121

$$5)\overline{170}$$
$$M = 34$$

$$5)\overline{228}$$
$$\sigma^2 = 45.6$$
$$\sigma = 6.75$$

Coef. $\sigma = 6.75 \div 34 = 0.20$

II. Grouped data

L_1–L_2	X	f	fX	x	x^2	fx^2
10–12	11	3	33	−3.72	13.8384	41.5152
12–14	13	15	195	−1.72	2.9584	44.3760
14–16	15	20	300	0.28	0.0784	1.5680
16–18	17	10	170	2.28	5.1984	51.9840
18–20	19	2	38	4.28	18.3184	36.6368

$$N = 50 \qquad 50)\overline{736}$$
$$M = 14.72$$

$$50)\overline{176.0800}$$
$$\sigma^2 = 3.5216$$
$$\sigma = 1.88$$

$$\sigma = \sqrt{\frac{\Sigma f x^2}{N}} = \sqrt{\frac{176.08}{50}} = 1.88$$

Coef. $\sigma = \sigma/M = 1.88 \div 14.72 = 0.13$, or 13 per cent

Short-cut method.—The calculation of the standard deviation may become tedious if actual deviations of the measures from the mean are taken, since the squares of these deviations are likely to run into several decimals. Hence the process commonly utilized involves selection of an arbitrary origin, as in calculating the arithmetic mean. Since deviations from the assumed origin are inaccurate, and the average of their squares has consequently been increased, a deduction for this inaccuracy must be made. The correction figure is, as may be demonstrated algebraically, the square of the correction figure (c) utilized in calculating the mean, or $\Sigma d \div N$. The basic procedure in thus calculating σ is illustrated in part I of Example 6·3. It will be noted that the symbol d is here used for the "uncentered" deviations from an assumed origin. As before noted, d' or D may be substituted to indicate the nature of the deviation.

In explanation of the procedure used in part II, it may be said that the assumed origin (R) is generally taken near the center of the dispersion, usually as the modal or largest frequency, because in this position it will reduce the deviations to small numbers. In summing the deviations it is necessary, of course, to multiply by the frequencies, and to *take account of the positive and negative signs*. The algebraic average of the deviations ($\Sigma fd \div N$) is called the correction (c), as previously explained. It will be noted that up to this point the process is the same as the so-called indirect or short-cut method of calculating the arithmetic mean. (See Example 5·1, page 91.)

In order to find the standard deviation, only one additional column is required. This is the fd^2 column, which accumulates the squares of the deviations. Each item in the column is readily obtained by multiplying the two preceding columns ($d \times fd = fd^2$), or by squaring the d column and multiplying by f. The two methods may be used advantageously to check each other:[1] thus for the first row, $d \times fd = (-4)(-12) = 48$,

---

[1] This check is a useful substitute for a more complicated device known as the Charlier check, which consists of adding 1 to each d to obtain D, then squaring and summing, taking account of frequencies. The result should obviously be

$$\Sigma fD^2 = \Sigma f(d+1)^2 = \Sigma fd^2 + 2\Sigma fd + N$$

<div align="center">

EXAMPLE 6·3

THE STANDARD DEVIATION—SHORT-CUT METHOD

</div>

Data: See Example 5·1, page 91.

I. Ungrouped data.   The assumed origin (R) is 30.

m or X	d	d^2
25	-5	25
$R = $ 30	0	0
33	3	9
37	7	49
45	15	225

$$\begin{array}{cc}
5)\underline{20} & 5)\underline{308} \\
c = 4 & 61.6 \\
R = 30 & c^2 = \underline{16} \\
M = \overline{34} & \sigma^2 = \overline{45.6} \\
& \sigma = 6.75 \\
\text{Coef. } \sigma = & 6.75 \div 34 \\
& = 0.20, \text{ or } 20 \text{ per cent}
\end{array}$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{61.6 - 4^2} = 6.75$$

II. Grouped data.[1]   The assumed origin (R) is 15.

L_1–L_2	m	f	$d = m - R$	fd	fd^2
10–12	11	3	-4	-12	48
12–14	13	15	-2	-30	60
14–16	$R = $ 15	20	0	0	0
16–18	17	10	2	20	40
18–20	19	2	4	8	32
		$N = 50$			

$$\begin{array}{cc}
50)\underline{-14} & 50)\underline{180} \\
c = -0.28 & d^2 = 3.6 \\
R = 15.00 & c^2 = 0.0784 \\
M = \overline{14.72} & \sigma^2 = \overline{3.5216} \\
& \sigma = 1.88 \\
\text{Coef. } \sigma = & 1.88 \div 14.72 \\
& = 0.13, \text{ or } 13 \text{ per cent}
\end{array}$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2} = \sqrt{3.6 - 0.28^2} = 1.88$$

or,

$$\sigma = \frac{\sqrt{N\Sigma d^2 - (\Sigma d)^2}}{N} = \frac{\sqrt{50 \times 180 - 14^2}}{50} = 1.88$$

For Example 6·3, part I₁:

$$\begin{array}{l}
D = -3; \quad -1; \quad 1; \quad 3; \quad 5 \\
f = 3; \quad 15; \quad 20; \quad 10; \quad 2. \quad \Sigma fD^2 = 202
\end{array}$$

Check: $\Sigma fd^2 + 2\Sigma fd + N = 180 - 2 \times 14 + 50 = 202$.

[1] If the deviations are expressed in units of class intervals (d_i) they become

or $f \times d^2 = 3 \times 16 = 48$. All items in the fd^2 column are necessarily positive. The column is totaled, and the uncorrected average d^2 is obtained by dividing by N. The correction figure squared (i.e., c^2) is subtracted from the average d^2, the result being the square of the standard deviation. The subtraction of the c^2 may be rationalized in that the error involved in the use of the arbitrary origin carries over as a square in each deviation, and, since squares are always positive, a subtraction must be made in order to correct for this error. In other words, the root-mean-square or standard deviation is a minimum when taken from the arithmetic mean, and when taken from any assumed mean must, therefore, be corrected by an appropriate subtraction. The validity of the rule may be logically established by algebraic analysis. The corrected average d^2 thus obtained is the square of the standard deviation and is often called the variance (V). Its square root may readily be taken to secure the standard deviation. The entire calculation may be carried through in units of class intervals, as is indicated in the footnote to Example 6·3.

The standard deviation, like other measures of dispersion, may be expressed as a coefficient, in order to facilitate comparison. Thus if the variability of a group of workers in respect to skill and in respect to earnings were to be compared, a coefficient of each measure would be required. For the given dispersion, this may be found as

$$\text{Coefficient } \sigma = \frac{\sigma}{M} = \frac{1.88}{14.72} = 0.13 = 13\%$$

Written as a percentage, such a ratio is called a coefficient of dispersion, a term applied to the percentage ratio of any measure of dispersion to a corresponding measure of central tendency.

Machine calculation of σ.—In machine calculation, or in dealing with very small numbers, probably the most convenient method of computing the standard deviation is to treat the

---

$-2; -1; 0; +1; +2$, and if the calculation is carried through on this basis, $c = -0.14$ and $\sigma_i = 0.94$. Hence, $\sigma = \sigma_i \times i = 0.94 \times 2 = 1.88$. Also $M = R + (c_i \times i) = 15 + 2(-0.14) = 14.72$.

original items as the uncentered deviations were treated in Example 6·3. In that case the m's or X's are squared, and the correction term, c, is replaced by M. Thus, in effect, the assumed mean is zero, from which the actual items are deviations. The method is expressed by the formula,

$$\sigma = \sqrt{\frac{\Sigma m^2}{N} - M^2}$$

where frequencies, if present, are taken account of in the summing. Or, in general, where X is any variable,

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - M^2} = \sqrt{\frac{\Sigma X^2 - NM^2}{N}}$$

where NM^2 is conveniently computed as $M\Sigma X$. The process is illustrated in Example 6·4. In machine calculation, of course, columns involving multiplication are not itemized, but are accumulated on the machine.

A further variation of the method, designed to avoid decimals, finds $N\Sigma x^2$ as follows:

$$N\Sigma x^2 = N\Sigma X^2 - (\Sigma X)^2$$

the square root of which is N times the standard deviation.[1]

Range of the standard deviation.—It will be noted that the standard deviation is somewhat larger than the average deviation of the same data, in general, by about 25 per cent. The standard deviation is larger because the process of squaring is analogous to weighting the items, and in effect the large deviations are weighted more heavily than the small deviations.

---

[1] The formulas utilized in machine calculation are obtained thus:

By definition,
$$\Sigma x^2 = \Sigma(X - M)^2 = \Sigma X^2 - 2M\Sigma X + NM^2$$
since $\Sigma X = NM,$ $\quad \Sigma x^2 = \Sigma X^2 - NM^2$
since $M^2 = (\Sigma X/N)^2,$ $\quad \Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/N$
and $\quad N\Sigma x^2 = N\Sigma X^2 - \Sigma X\Sigma X$

<p style="text-align:center">EXAMPLE 6·4</p>

<p style="text-align:center">THE STANDARD DEVIATION—MACHINE METHOD</p>

Data: See Example 5·1, page 91.

I. Ungrouped data

X	X^2		X^2
25	625		625
30	900		900
33	1,089	or	1,089
37	1,369		1,369
45	2,025		2,025

$$5)\overline{170} \qquad 5)\overline{6,008} \qquad \Sigma X^2 = 6,008$$

$$M = 34 \qquad 1,201.6 \qquad M\Sigma X = 5,780 = NM^2$$

$$M^2 = \overline{1,156.0} \qquad 5)\overline{228}$$

$$\sigma^2 = \quad 45.6 \qquad \sigma^2 = \quad 45.6$$

$$\sigma = \quad 6.75 \qquad \sigma = \quad 6.75$$

II. Grouped data

X	f	fX	fX^2		fX^2
11	3	33	363		363
13	15	195	2,535		2,535
15	20	300	4,500	or	4,500
17	10	170	2,890		2,890
19	2	38	722		722

$$N = 50 \qquad 736 \qquad 50)\overline{11,010} \qquad 11,010$$

$$M = \overline{14.72} \qquad \overline{220.2} \qquad M\Sigma X = 10,833.92 = NM^2$$

$$M^2 = 216.6784 \qquad 50)\overline{176.08}$$

$$\sigma^2 = \quad 3.5216 \qquad \sigma^2 = \quad 3.5216$$

$$\sigma = \quad 1.88 \qquad \sigma = \quad 1.88$$

or

$$N\Sigma x^2 = N\Sigma X^2 - \Sigma X \Sigma X$$

$$= 50 \times 11,010 - 736 \times 736 = 550,500 - 541,696 = 8,804$$

$$\sigma = \frac{\sqrt{N\Sigma x^2}}{N} = \frac{93.8296}{50} = 1.8766$$

By reference to Example 6·3, part II, it will be seen that the standard deviation is 1.88 as measured from the mean, 14.72. That is, the standard deviation implies a range of 14.72 ± 1.88 or from 12.84 to 16.60.   As in the case of the average deviation this range does not include the extreme variations but represents the variability with the large and small variations averaged by the use of a quadratic mean.   The entire variability of the tabulation lies within the range of two or three standard deviations below and above the mean.   That is, 14.72 ± (2 × 1.88) indicates a spread from 10.96 to 18.48, which includes most of the items in the tabulation, while 14.72 ± (3 × 1.88) includes a spread from 9.08 to 20.36, which extends beyond the limits (10 and 20) of the tabulation.

Dispersion measured from the median.—The median is sometimes used as the origin from which to measure dispersions. When it is so utilized, the average deviation is appropriate, since *the average deviation taken from the median is a minimum*; that is, it is as small as or smaller than the average deviation taken from any other origin.[1]   The calculation is not different from that in which the average deviation is taken from the mean, except that deviations are equal to $m - Md$.

Quartile deviation and percentiles.—Another measure of dispersion connected with the median is that known as *quartile deviation*.   This measure, like the median, discounts the magnitude of the extreme items and is, therefore, indicated for data to which the median is adaptable.   It is especially useful when the importance or accuracy of extremely large or small items is in question, and with " open-class " distributions, to be discussed later.

Measurement of quartile deviation necessitates a division of the distribution into 4 parts, or quartiles, in the same way that the median divides it into 2 parts.   That is, the first quartile

---

[1] This statement is subject to some qualification so far as the average deviation in grouped data is concerned, on account of the assumption that the mid-points or class measures fairly represent the items in each class.   In practice, because of the error introduced by reference to these mid-points as class measures, the average deviation of grouped data from the median may be found to be larger than that when the mean is regarded as the origin.   Thus in the case of the data used as illustrative in Example 6·1, II, the AD_M is found as 1.478, while $AD_{Md} = 1.484$.

(Q_1) may be described as the median of the first half of the distribution, and the third quartile (Q_3) as the median of the second half of the distribution.[1]   Thus the Q_1 position is the space numbered $N \div 4$ in the array or tabulation.   Likewise, the third quartile position is the space numbered $3N \div 4$.   It is obvious that the second quartile (Q_2) is identical with the median.

The calculation of the quartile deviation is illustrated in Example 6·5.   Whether the data are ungrouped or grouped, the first step is to find Q_1 and Q_3 by the methods just indicated. Then, the quartile deviation (QD) is obtained as

$$QD = \frac{Q_3 - Q_1}{2}$$

Sometimes it is convenient to make a still further division into *percentiles*, which divide the distribution into hundredths. Thus, the median is the fiftieth percentile, the first quartile the twenty-fifth percentile, and the third quartile the seventy-fifth percentile.   In Example 6·5 the formula for the quartiles is given in a general form applicable to any percentile (P), as follows:

$$P = \left(\frac{FN - \Sigma_1}{f} \times i\right) + L_1$$

where F is the percentile expressed as a decimal fraction as 0.50 for the median or 0.10 for the tenth percentile and Σ_1 and L_1 refer to the class in which the percentile appears.   The formula is applicable to the determination of any percentile as illustrated in Example 6·5, where FN is written $xN/100$.

The coefficient of quartile deviation.—The coefficient of quartile deviation, which is useful when different types of distributions are compared, is generally found by comparison

---

[1] Just as the quartiles divide a series into 4 parts each containing theoretically an equal number of items, so the quintiles divide it into 5 parts, the deciles into 10 parts, and the percentiles into 100 parts.   It should also be noted that these terms may be used to designate a range limited by the measure thus designated.   For example, the statement that a given wage falls in the upper quartile means that it is above the third quartile.   Similarly, an item falling in the second quintile means that it is between the first and second quintile magnitudes.

<div align="center">

EXAMPLE 6·5

QUARTILES AND PERCENTILES

</div>

Data: See Example 5·1, page 91.

  I. Ungrouped data

Items arrayed	25	30	33	37	45	55
Spaces between items (numbered)	1	2	3	4	5	

Quartiles:

    Q_1 at space $\frac{1}{4}N$ = space $(\frac{1}{4} \times 6)$ = space $1.5 = 30$.

    Q_3 at space $\frac{3}{4}N$ = space $(\frac{3}{4} \times 6)$ = space $4.5 = 45$.

    (If N is odd, the quartile may be regarded as the item nearest the indicated position.)

Quartile deviation:

$$QD = \frac{(Q_3 - Q_1)}{2} = \frac{(45 - 30)}{2} = 7.5.$$

  Coef. $QD = \frac{(Q_3 - Q_1)}{(Q_3 + Q_1)} = \frac{15}{75} = 0.20.$

Percentiles:

  P_x (approximate) at $\frac{xN}{100}$ space, or nearest item, i.e.,

  P_{60} at space $\left(\frac{60}{100} \times 6\right)$ = space 3.6; $P_{60} = 37$ (approximate).

  II. Grouped data

Class limits		Frequencies	Cumulatives	
L_1	L_2	f	Σ_1	Σ_2
10	12	3	0	3
12	14	15	3	18 (Q_1 class)
14	16	20	18	38 (Q_3 class)
16	18	10	38	48 (P_{90} class)
18	20	2	48	50
		50		

Quartiles:

$$Q_1 = \left(\frac{(N \div 4) - \Sigma_1}{f}\right) \times i + L_1 = \frac{12.5 - 3}{15} \times 2 + 12 = 13.27.$$

$$Q_3 = \left(\frac{(3N \div 4) - \Sigma_1}{f}\right) \times i + L_1 = \frac{37.5 - 18}{20} \times 2 + 14 = 15.95.$$

Quartile deviation:

  $QD = (Q_3 - Q_1) \div 2 = (15.95 - 13.27) \div 2 = 1.34.$

  Coef. $QD = (Q_3 - Q_1) \div (Q_3 + Q_1) = 0.092.$

Percentiles:

$$P_x = \left(\frac{FN - \Sigma_1}{f}\right)(i) + L_1; \text{e.g., } P_{90} = \left(\frac{(45 - 38)}{10}\right)(2) + 16 = 17.40$$

with the so-called mid-quartile measure, that is, the value half way between the first and third quartiles.   In normal distributions this would obviously be identical with the median.   If the mid-quartile measure (MQ) is calculated, it is expressed by the formula,

$$MQ = \frac{Q_3 + Q_1}{2}$$

It is not, however, necessary to calculate this measure, which is of no particular interest in itself.   Instead, its algebraic equivalent may be substituted in the formula to secure the coefficient of quartile deviation, as follows:

$$\text{Coef. } QD = QD \div MQ = \frac{Q_3 - Q_1}{2} \div \frac{Q_3 + Q_1}{2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Hence, the coefficient of quartile deviation may be calculated directly from the first and third quartiles by taking the ratio of their difference to their sum.

Percentiles in an array.—The interpolation of percentiles in an array of untabulated data is sufficiently accurate, as a rule, if the percentile magnitude is rounded to either the nearest item or the average of two adjacent items (see Example 6·5, part I).   If, however, more accurate interpolation is required, this result may be accomplished on the assumption that the items in the array represent a tabulation having as its successive class limits the items and the mid-points between the items with open classes at each end.   The frequencies in each class may be taken as unity, in which case N is twice the actual number of items, since there are two classes for each item.   The magnitude, FN, may then be interpolated in the cumulatives, as in any problem with grouped data (see Appendix, page 515).

It is sometimes desirable to calculate the position of a given measure in a distribution, that is, to find FN and F from a given percentile magnitude, P.   This may readily be done by reversing the interpolation equation previously given so as to express the value of FN as follows:

$$FN = \left(\frac{P - L_1}{i} \times f\right) + \Sigma_1$$

The fractional position is

$$FN \div N = F$$

To illustrate, in Example 6·6, the proportion of incomes below $800 may be found as

$$100F = \frac{800 - 750}{250} \times 14.90 + 31.64$$

$$= 34.62$$

that is, it may be roughly estimated that approximately 35 out of 100, or 35 per cent, of all incomes were below $800.
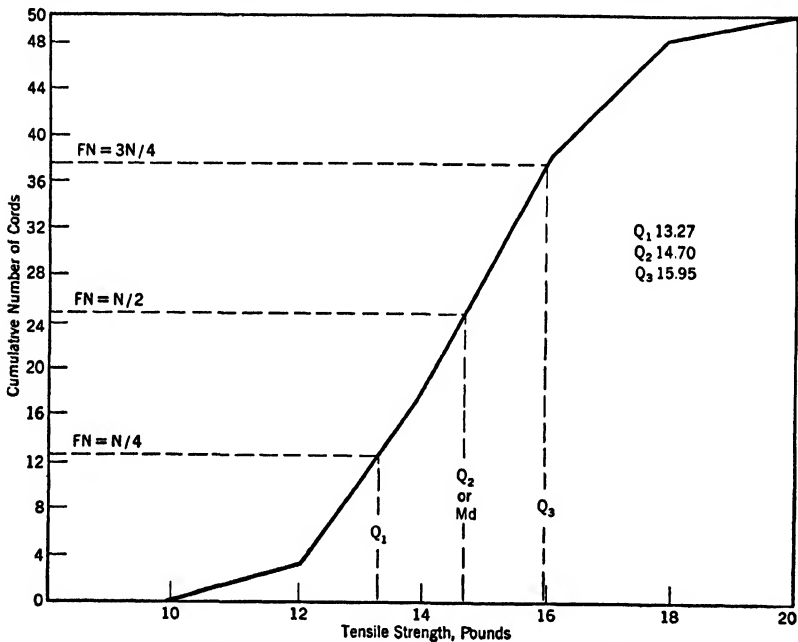
FIG. 6·1.—Graphic Estimation of the Quartiles. Data: See Example 6·5, part II.

Graphic interpolation.—The quartiles or other percentiles may often be estimated with sufficient accuracy by means of an ogive of the distribution, as illustrated in Fig. 6·1. If the median, for example, is required, $N/2$ is located on the vertical scale of " less than " cumulatives. As explained in the preceding chapter, a point on the ogive directly to the right is then

located, and the median is read on the horizontal scale perpendicularly below. In the same way other percentiles may be read. Interpolation may be reversed, to estimate the number of workers below or above a given wage.

The use of the median as a measure of central tendency together with the quartile deviation as a measure of dispersion will be found very convenient in connection with a certain type of distribution frequently found in statistical reports. For example, the distribution of income by income groups is very commonly given in classes of variable size and with an "open class " at one or both extremes of the distribution.[1] An income distribution may begin with the class "under $250," and end with the class "$100,000 and over" (see Example 6·6). Or an age distribution may end with the class "80 or more." Unless recourse can be had to the original data, such distributions cannot be typified accurately by their mean, nor can their dispersion be measured by their average or standard deviations, but the median and quartiles may be calculated by the method described. When the class interval is variable, care must be taken to use the particular interval applicable to each calculation. If the distribution is reasonably normal, the standard deviation may be approximated as $\sigma = QD \div 0.6745$.

These various measures of dispersion, the average deviation, standard deviation, and quartile deviation, are useful in many connections, as will be apparent from subsequent discussions. They measure the tendency of the data to vary from the average of the group as a whole. They thereby give added meaning

[1] Broad income distributions generally appear to be logarithmic normals, that is, they seem to take approximately a normal form when the frequencies are plotted against the logarithms of the class limits and measures. A more careful study, however, usually indicates that the incomes in the higher classes extend far beyond the normal suggested by the lower ranges. This departure from normality is apparently due to the fact that there are two distinct types of income, namely, those that are attributable to direct personal services, and those that are attributable to capital ownership, though any individual income may be a combination of the two types. The abnormality may be visualized if the distribution is plotted on double-logarithmic paper. A log normal distribution should then appear as a parabola (an inverted U) but in fact the larger incomes then approximate a straight line. Pareto first noted this abnormality, though he plotted the upper "tail" of the distribution as a "more than" cumulative, which on such a chart also approximates a straight line when the curve itself does so.

EXAMPLE 6·6

LOCATION OF QUARTILES IN IRREGULARLY GROUPED DATA

Data: Distribution of families and single individuals by percentage of national income received, United States, 1935–1936 ("Consumer Incomes in the United States," National Resources Committee).

Income level	Frequencies		Cumulative, $\Sigma_2\%$
	Number	$f\%$	
Under $250.....................	2,123,534	5.38	5.38
$250–$500......................	4,587,377	11.63	17.01
$500–$750......................	5,771,960	14.63	31.64
$750–$1,000...................	5,876,078	14.90	46.54
$1,000–$1,250.................	4,990,995	12.65	59.19
$1,250–$1,500.................	3,743,428	9.49	68.68
$1,500–$1,750.................	2,889,904	7.32	76.00
$1,750–$2,000.................	2,296,022	5.82	81.82
$2,000–$2,250.................	1,704,535	4.32	86.14
$2,250–$2,500.................	1,254,076	3.18	89.32
$2,500–$3,000.................	1,475,474	3.74	93.06
$3,000–$3,500.................	851,919	2.16	95.22
$3,500–$4,000.................	502,159	1.27	96.49
$4,000–$4,500.................	286,053	0.72	97.21
$4,500–$5,000.................	178,138	0.45	97.66
$5,000–$7,500.................	380,266	0.96	98.62
$7,500–$10,000................	215,642	0.55	99.17
$10,000–$15,000..............	152,682	0.39	99.56
$15,000–$20,000..............	67,923	0.17	99.73
$20,000–$25,000..............	39,825	0.10	99.83
$25,000–$30,000..............	25,583	0.06	99.89
$30,000–$40,000..............	17,959	0.05	99.94
$40,000–$50,000..............	8,340	0.02	99.96
$50,000–$100,000.............	13,041	0.03	99.99
$100,000 and over............	5,387	0.01	100.00
All levels.................	39,458,300	100.00	

General formula for any percentile (P):

$$P = \frac{FN - \Sigma_1}{f} i + L_1$$

where F is the percentile expressed as a decimal fraction, and Σ_1 and L_1 refer to the class in which that percentile occurs.

Solution for first, second, and third quartiles (note that $\Sigma_1\%$ may be read in the table as $\Sigma_2\%$ of the preceding class):

$$Q_1 = \frac{25 - 17.01}{14.63} \times 250 + 500 = 637$$

$$Q_2 = \frac{50 - 46.54}{12.65} \times 250 + 1,000 = 1,068$$

$$Q_3 = \frac{75 - 68.68}{7.32} \times 250 + 1,500 = 1,716$$

to the average at the same time that they facilitate comparisons of this type of variability in different distributions. Many illustrations of such analysis might be cited. In one of these, for instance, the British Fatigue Research Board has found it possible to appraise comparative levels of unrest and restriction of output among workmen by noting time-to-time changes in the range of variation about the mean of production. In this analysis, the output of individual workers is classified as to quantity, and the numbers of employees in each such class are carefully noted. In effect, there is created a frequency distribution in which classes represent various levels of productivity and frequencies are the numbers of workers in each class. Average productivity is carefully noted, as is the range of variability about this average. For the latter purpose, the standard deviation is used, and a coefficient of variability, the coefficient of standard deviation described in preceding paragraphs, is calculated. These measures for the distributions representing various months are compared, and when the coefficient of variability for any month is notably reduced, that fact is regarded as a definite indication of intentional restriction of output.

Inaccuracies in grouped data.—It should be observed that the measurement of central tendency and dispersion based on grouped data involves certain minor inaccuracies and yields approximations only. Though the mean [1] is not seriously

---

[1] If the mean calculated from grouped data is compared with the mean of the crude data from which the grouping is derived, it may be found to vary a little. The difference thus found is known as the error of grouping or of tabulation. Though the error may sometimes be zero and occasionally may approach half a class interval $(i/2)$ in magnitude, its average size in many like problems may be estimated as follows:

$$\text{Average grouping error of } M, (GE_M) = \frac{0.23i}{\sqrt{N}}$$

The average error thus obtained is useful as indicating the danger of inaccuracy occasioned by tabulation. To illustrate, in Example 6·2 (part II), page 114, the average error of M due to tabulation is

$$GE_M = (0.23 \times 2) \div \sqrt{50} = 0.065$$

hence the mean as calculated, 14.72, is probably too large or too small by as much as 0.07. The error of grouping may be expected to appear also in the standard deviation. However, it should be remembered that the error may be reduced by a careful choice of class limits.

affected, the median and other percentiles, interpolated on the assumption of a regular distribution of the items in each class, may be somewhat biased.[1]

The most important inaccuracy attributable to the process of grouping data is found in the standard deviation. In this case, there are two partially offsetting biases, which are designated as the "squares" and "slope" biases, respectively. By the squares bias is meant the fact that, in any class, fx^2 (i.e., $f[X - M]^2$) is not truly representative of the deviations of the items assumed to be scattered regularly throughout the class. For example, suppose that $X - M$ is 5 and the actual individual deviations which it represents are 2, 4, 6, and 8. Then fx^2 or 4×5^2 is 100, while the sum of 2^2 plus 4^2 plus 6^2 plus 8^2 is 120. The square of the class mark thus underestimates the actual sum of squared deviations.

The second or "slope" bias, however, has an even stronger upward pull. It arises from the tendency of items in a class to reflect the position of the mode of the distribution. Obviously, the items in any class tend to cluster on the side of the class nearer to the mode. Described in terms of a graphic representation, the frequencies in the class slope upward toward the mode, instead of being horizontal throughout the class.

Possible corrections.—If the distribution has what is called "high order contact," that is, if it has many decreasing classes at each extreme, then the adjustment known as "Sheppard's correction" will make proper allowance for the biases of squares and slope. It involves subtraction of $i^2/12$ from the variance as ordinarily computed from grouped data. The correction may be applied to the data of Part II, Example 6·4, page 119, which have fairly high order contact, as follows:

$$\sigma_c^2 = \frac{\Sigma x^2}{N} - \frac{i^2}{12} = 3.5216 - \frac{4}{12} = 3.1883$$

$$\sigma_c = 1.786$$

---

[1] By parabolic interpolation, as determined by frequencies preceding (f_{-1}) and following (f_{+1}), a percentile indicated by the fraction of FN/N is

$$P = L_1 + i\left(\pm\sqrt{\frac{4(FN - \Sigma_1)}{f_{+1} - f_{-1}} + \left(\frac{f - f_{+1} + f_{-1}}{2(f_{+1} - f_{-1})}\right)^2} - \frac{4f - f_{+1} + f_{-1}}{2(f_{+1} - f_{-1})}\right)$$

If, however, the distribution under consideration does not have high order contact, then a more complicated adjustment is necessary. If the distribution may reasonably be regarded as continuous, rather than discrete, then the correction may be applied to the summed squares, as follows:

$$\Sigma f X_c^2 = \Sigma f X^2 - [N - \tfrac{4}{5}(f_a + f_z)]\frac{i^2}{12}$$

and the corrected variance may be found as,

$$\sigma_c^2 = \frac{1}{N}(\Sigma f X_c^2 - M\Sigma f X)$$

where f_a and f_z are the frequencies of the first and last classes in the distribution.[1] As illustrated by the data of the preceding paragraph, the correction may be applied as follows:

$$\Sigma f X_c^2 = 11{,}010 - [50 - \tfrac{4}{5}(3 + 2)]\tfrac{4}{12} = 10{,}994.67$$

$$\sigma_c^2 = \tfrac{1}{50}[10{,}994.67 - (14.72 \times 736)] = 3.215$$

$$\sigma_c = 1.793$$

Like the standard deviation, the average deviation computed from grouped data is also subject to two partially offsetting biases. The first of these biases—negative in its tendency—arises in the class in which the origin lies, and is called

---

[1] It may be noted that, if the distribution is a continuous one and has high order contact, this correction reduces to Sheppard's. It is considerably more detailed and flexible, however, as it is based on parabolic smoothing, and it may be applied whether high order contact is present or not. Where distributions under consideration are discrete rather than continuous, the correction formula requires some modification involving the addition of a k factor, in which k stands for the number of integral subclasses in each class. Thus, if each class includes five integers or subclasses, then $k = 5$. The correction formula for discrete distributions is

$$\Sigma f X_c^2 = \Sigma f X^2 - \left[N - \frac{4}{5}(f_a + f_z)\frac{k^2 + 1}{k^2}\right]\left[\frac{i^2}{12} \times \frac{k^2 - 1}{k^2}\right]$$

$$\sigma_c^2 = \frac{1}{N}(\Sigma f X_c^2 - M\Sigma f X)$$

the bias of the origin class. The second—a positive bias—is similar to the *slope* bias of the standard deviation.[1]

The bias of the origin class is at its maximum when the origin is at the class mark. In that case all the deviations of items in this class are lost when m is taken to represent them. If the origin is close to a class limit, however, the correction may become negligible. In general, however, the slope bias is likely to predominate. But in any ordinary case, it seems hardly worth while to apply correction formulas in computing the average deviation, inasmuch as it is a convenient and approximate rather than a scientific and exact measure of dispersion. Moreover, to apply one correction without the other may make the result worse rather than better.

Mention should be made of certain graphic methods which are useful in obtaining improved estimates of the median or other percentiles, as well as in fitting curves to normal or logarithmic normal distributions. (See Chapter VII for a description of such distributions.) These methods make use of so-called *arithmetic probability paper*. The vertical scale of this paper is ruled in such a way as to reduce the ogive of a normal distribution (and approximately a binomial distribution) to a

---

[1] The bias of the origin class may be offset by *adding* a correction (C_1) to $\Sigma \mid d \mid$ as ordinarily computed, as follows:

$$C_1 = \frac{f_r}{4i} (i - 2 \mid d_r \mid)^2$$

where f_r is the frequency of the origin class, and d_r is the deviation of m or X of the same class, that is, $\mid X - R \mid$.

The correction (C_2) to be *subtracted* for slope in the non-origin classes may be taken as

$$C_2 = \frac{i}{24} (f_{-1} + f_{+1} + 2f_r)$$

where f_{-1} and f_{+1} are the frequencies preceding and following f_r (at next smaller and larger X's), respectively. It should be noted that C_1 is added and C_2 is subtracted.

As applied to the data of Example 6·1 (part II), page 110, the corrected average deviation becomes

$$AD_c = [\Sigma f \mid d \mid + C_1 - C_2] \div N$$

$$= \left[73.92 + \frac{20}{4 \times 2} (2 - 2 \times 0.28)^2 - \frac{2}{24} (15 + 10 + 2 \times 20) \right] \div 50$$

$$= (73.92 + 5.184 - 5.417) \div 50 = 1.4737$$

straight line (cf. Figs. 6·2 and 6·3).   The frequencies of the
distribution must, however, be reduced to percentages of N to
be adapted to the printed ogive scale, which approaches 0 and
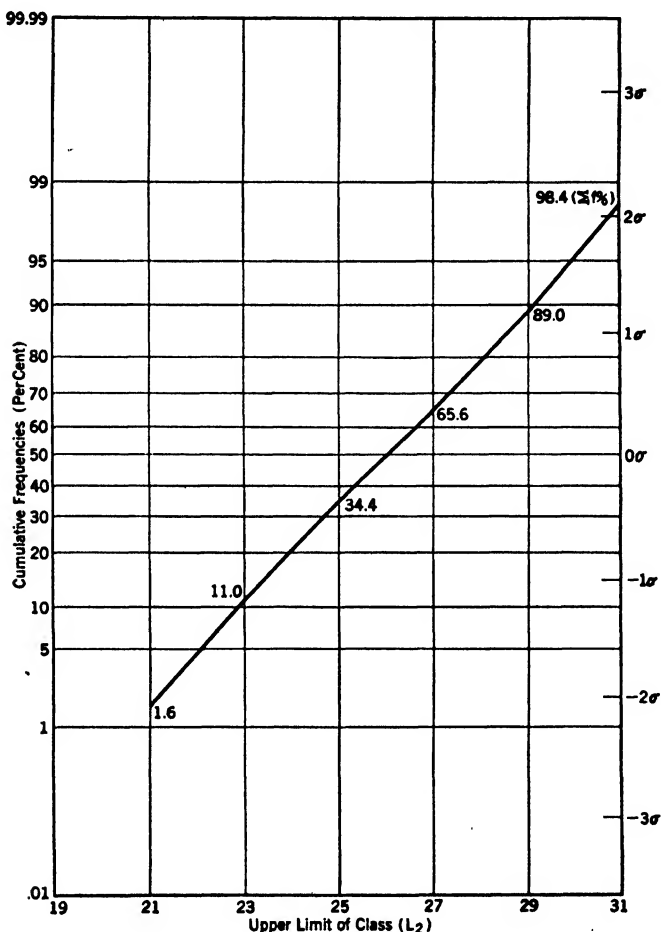
Fɪɢ. 6·2.—Probability Cumulative Curve of a Seven Class Binomial Distribution
Having the Following Percentage Frequencies: 1.6; 9.4; 23.4; 31.2; 23.4; 9.4; 1.6.

100 as its limits.   By drawing the cumulative curve on this
type of paper, and interpolating the quartiles, deciles, or other
percentiles in the same manner as on ordinary cumulative charts
(see Fig. 6·1), it is frequently possible to obtain more accurate

estimates than those obtained by the ordinary linear interpolation formulas previously described. Probability paper is also useful in subdividing open classes at the extremes of a distribu-
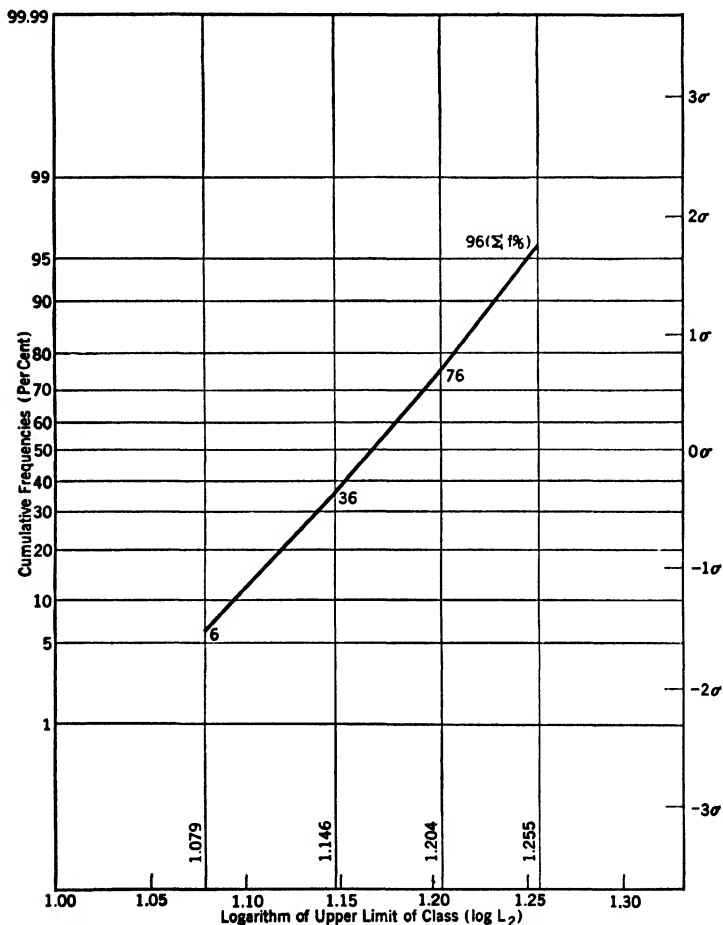
FIG. 6·3.—Logarithmic Probability Cumulative Curve of Data of Example 6·3, part II, page 116 ($\Sigma f\%$ plotted against log L_2).

tion. This result may be secured merely by extending the curve by inspection, and reading the cumulative frequencies against successive upper limits of the classes required. The frequencies may then be obtained by subtraction.

Probability paper is sometimes printed with a logarithmic instead of an arithmetic scale on the X axis. Since many distributions approximate the so-called logarithmic type, the use of this scale in such cases eliminates the skewness, and the cumulative curve again appears as a straight line. However, the same effect may be similarly obtained on arithmetic probability paper by plotting the cumulative curve against the logarithms of the class limits (see Fig. 6·3). When skewness is irregular it may sometimes be eliminated by plotting the cumulative curve against the logarithms of the class limits plus or minus some constant. Such adjustments frequently involve complex calculations and are beyond the range of this discussion. In the Appendix, the fitting of a normal and logarithmic normal curve to data by the use of probability paper is discussed (see Appendix, page 517).

Although probability scales are advantageous in estimating and removing the inaccuracies due to grouping, the same purposes may also be served by numerous formulas which are available in advanced textbooks. It is generally preferable, if greater precision is required than is obtained by the usual methods, to resort to the original data, and either to make the calculations on the basis of these ungrouped figures or else to tabulate the data in classes sufficiently small so that errors become negligible.

Measures of dispersion compared.—The several measures of central tendency and related measures of dispersion described in preceding pages may well be briefly summarized and compared. In summary form, these relationships are:

MEASURE OF CENTRAL TENDENCY	RELATED MEASURE OF DISPERSION
(1) Arithmetic mean (M) $$M = \Sigma m/N \text{ or } \Sigma X/N$$	Average deviation (AD) $$AD = \Sigma\,|\,d\,| \div N$$
(2) Arithmetic mean (M) $$M = \Sigma m/N$$	Standard deviation (σ) $$\sigma = \sqrt{\Sigma x^2 \div N}$$
(3) Geometric mean (GM) $$\log GM = \Sigma \log m \div N$$	Standard deviation ratio (σ_r) $$\log \sigma_r = \sigma \text{ of } \log m\text{'s}$$
(4) Median (Md) $$Md = [(N/2 - \Sigma_1)/f]i + L_1$$	Average deviation (AD) $$AD = \Sigma\,|\,d\,| \div N$$

Measure of Central Tendency	Related Measure of Dispersion
(5) Mid-quartile measure (MQ)	Quartile deviation (QD)
$$MQ = \frac{(Q_3 + Q_1)}{2}$$	$$QD = (Q_3 - Q_1) \div 2$$
(6) Mode (Mo)	Average deviation (AD)
$Mo = [d_1 \div (d_1 + d_2)]i + L_1$	occasionally used
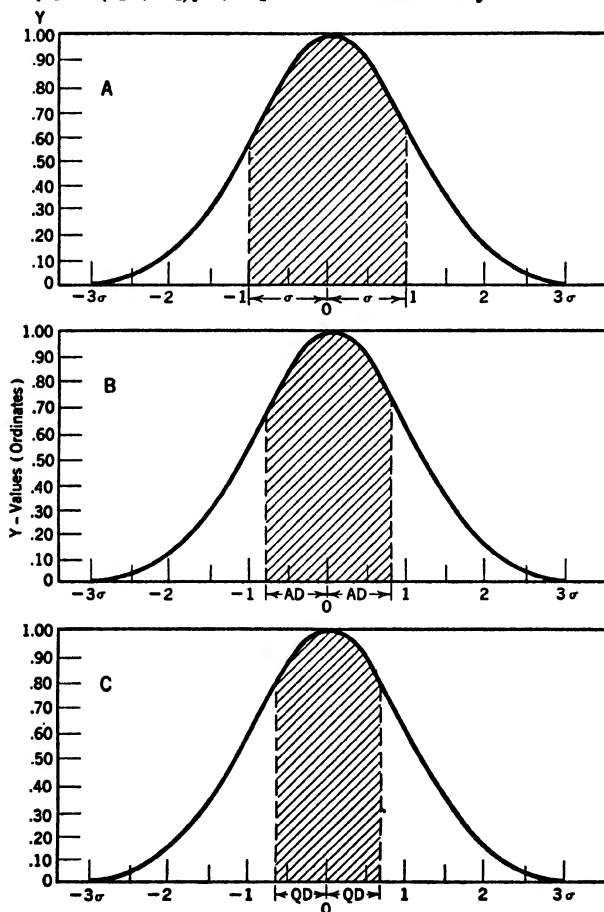
Fig. 6·4.—Comparison of Standard Deviation (A), Average Deviation (B), and Quartile Deviation (C) in a Normal Distribution. Data: $\sigma = 1$; $AD = 0.7979$; $QD = 0.6745$.

The most commonly used measures of dispersion are graphically compared in Fig. 6·4. In the figure, the Y scale represents proportions of the maximum ordinate, Y_o, at the mean. It may

be noted, with respect to the six methods of measurement, that (1) method 1 is suitable for use with ordinary problems not calling for extended mathematical analysis; (2) method 2 is a standard mathematical procedure; (3) method 3 is sometimes similarly used with logarithmic normal distributions; (4) method 4 is an alternate for (1) in which the absolute deviations are minimized; (5) method 5 with the median included is frequently useful with open-class distributions, or in cases where little emphasis is to be given to extreme magnitudes; (6) method 6 puts less stress on such extreme items.

The measurement of skewness (Sk).— The normal distribution, as has been noted, is symmetrical. As a result, mean, median, and mode coincide. When a distribution is featured by asymmetry, when it is not

Fig. 6·5.—A Positively Skewed Distribution. Frequency distribution of duration of business cycles in the United States, 1796–1923.

symmetrical, it is said to be *skewed*. One means of noting and measuring the lack of symmetry involves a comparison of the relative position of mean, mode, and median. In a skewed distribution, the mean and mode will be most distant, the mean and median less widely separated (see footnote, page 103).

These characteristics of skewed distributions may be seen in Figs. 6·5 and 6·6. Figure 6·5 represents a positively skewed distribution; Fig. 6·6, a negatively skewed one. It may be noted that, as a rule, if the mean exceeds the mode, the distribution is positively skewed, whereas if the mode exceeds the mean it is negatively skewed.

There are several commonly used methods of measuring skewness. Because of the fact that skewness appears clearly in the separation of the mean and the mode, some measures refer to these statistics. Thus, for the distribution shown in

Fig. 6·5 (where $M = 4.031$, $Mo = 3.046$, and $\sigma = 1.686$), the skewness may be measured as

$$Sk = \frac{M - Mo}{\sigma} = \frac{4.031 - 3.046}{1.686} = 0.58$$

However, the mode is so infrequently used for other purposes, and is for that reason so seldom available, that effort is usually made to secure an approximate appraisal of skewness from other, more readily available measures of the distribution. Two
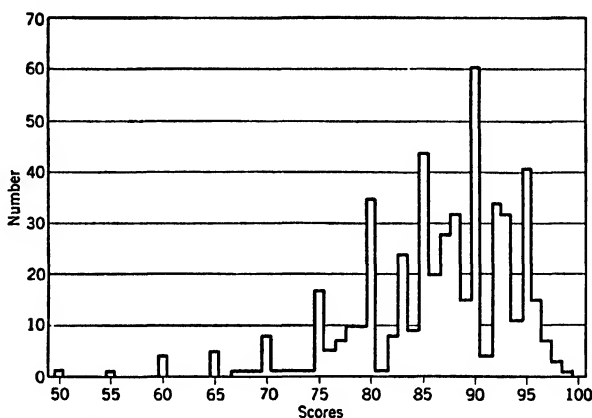
FIG. 6·6.—A Negatively Skewed Distribution of Certain Scholarship Grades.

methods of procedure are sufficiently widely used to justify attention here. They are applicable, however, only to distributions where skewness is not acute. The first is based on the fact that, in most cases, the difference between the mean and median is about one-third that between the mean and mode. Hence, it may be said that

$$Sk = \frac{3(M - Md)}{\sigma}$$

For the data of Fig. 6·5 (where $M = 4.031$, $Md = 3.672$, and $\sigma = 1.686$), the measure is

$$Sk = \frac{3(4.031 - 3.672)}{1.686} = 0.64$$

The second commonly used method, suggested by Professor Bowley, describes a measure of skewness which varies between -1 and $+1$. It refers only to the quartiles of the distribution and finds the degree of skewness as

$$Sk = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

For the distribution shown in Fig. 6·5 (where $Q_1 = 2.805$, $Q_2 = 3.672$, and $Q_3 = 5.157$), the degree of skewness is, therefore,

$$Sk = \frac{5.157 + 2.805 - 7.344}{5.157 - 2.805} = 0.26$$

It will be apparent that this measure is not at all comparable to either of the others described above. When reference is made only to the quartiles, a measure as great as 0.10 shows considerable skewness, and a value in excess of 0.30 would be found only in cases of unusual skewness.[1]

---

[1] Comparisons of curves representing frequency distributions with the normal curve frequently refer to their relative roundness or flatness, and this characteristic is known as *kurtosis*. A measure of this curvature for the normal curve is expressed by the equation

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\Sigma x^4 \div N}{(\Sigma x^2 \div N)^2} = 3$$

where x refers to deviations from the mean of the distribution, and N is, as usual, the number of items.

For the normal curve, as indicated by the equation, this ratio is equal to 3, and when this condition characterizes any distribution, it is said to be *mesokurtic*, or equal, in this respect, to the curve of the normal distribution. When the ratio for the given distribution is less than 3, the curve for this distribution shows greater flatness than that of the normal, and the curve is said to be *platykurtic*. On the other hand, when the ratio is greater than 3, there is a stronger tendency to peak in the curve, and it is said to be *leptokurtic*.

Kelley has suggested as a simple formula for kurtosis (P_{75} is the seventy-fifth percentile, or third quartile, etc.):

$$Ku = \frac{P_{75} - P_{25}}{P_{90} - P_{10}}$$

A distribution is considered platykurtic if $Ku > 0.26315$.

READINGS

(Also see special and general references, pages 591 and 597.)

DAVIES, G. R., "The Analysis of Frequency Distributions," *Journal of the American Statistical Association*, 24 (168), December, 1929, pp. 349–366.

FERGER, WIRTH F., "The Nature and Use of the Harmonic Mean," *Journal of the American Statistical Association*, 26 (173), March, 1931, pp. 36–40.

HOJO, TOKISHIGE, "Distribution of the median, quartiles, and interquartile distance in samples from a normal population," *Biometrica*, 23 (3–4), December, 1931, pp. 315–360.

MAVERICK, LEWIS A., "Graphic Presentation of Standard Deviation," *Journal of the American Statistical Association*, 27 (179), September, 1932, pp. 287–297.

SHOOK, B. L., "A Synopsis of Elementary Mathematical Statistics," *Annals of Mathematical Statistics*, 1 (3), August, 1930, pp. 224–259.

WALKER, HELEN M., *Mathematics Essential for Elementary Statistics*, New York, Henry Holt & Co., 1934.

YANG, SIMON, "On Partition Values," *Journal of the American Statistical Association*, 28 (182), June, 1933, pp. 184–191.

ZIZEK, F., *Statistical Averages*, New York, Henry Holt & Co., 1913.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Measure dispersion in the following distributions by means of AD, coef. AD, σ, coef. σ, QD, and coef. QD.

(a)		(b)		(c)		(d)		(e)	
X	f	X	f	X	f	X	f	X	f
2	3	4	2	1	1	2	1	2	1
3	5	6	4	2	3	4	2	4	4
4	6	8	6	3	5	6	5	6	6
5	4	10	5	4	2	8	3	8	4
6	2	12	3	5	1	10	1	10	1

2. Measure dispersion in the following distributions by securing AD, σ, coef. σ, QD, and coef. QD.

(a)		(b)		(c)		(d)		(e)		(f)		(g)		(h)		(i)	
X	f	X	f	X	f	X	f	X	f	X	f	X	f	X	f	X	f
5	1	6	2	3	20	2	10	8	30	4	4	10	3	4	4	6	2
15	4	18	4	5	50	4	40	10	60	8	7	20	7	8	6	10	5
25	3	30	3	7	40	6	50	12	50	12	5	30	6	12	5	14	6
35	2	42	1	9	10	8	20	14	20	16	3	40	3	16	3	18	4
										20	1	50	1	20	2	22	3

3. Measure dispersion in the following distributions by finding AD, σ, and QD.

(a)		(b)		(c)		(d)		(e)		(f)	
X	f	X	f	X	f	X	f	X	f	X	f
2	3	3	1	3	1	2	2	3	1	6	2
4	9	5	6	5	7	4	5	4	14	8	12
6	10	7	7	7	11	6	7	5	25	10	24
8	8	9	7	9	9	8	6	6	27	12	25
10	6	11	4	11	6	10	5	7	18	14	17
12	3 -	13	3	13	4	12	3	8	9	16	10
14	1	15	2	15	2	14	2	9	4	18	7
								10	2	20	3

4. Find the mean and σ of the following distributions:

(a)		(b)		(c)		(d)		(e)	
X	f	X	f	X	f	X	f	X	f
20	1	20	1	2	4	1	1	10	2
40	5	40	4	4	7	2	4	12	5
60	3	60	6	6	5	3	5	14	6
80	1	80	4	8	3	4	3	16	4
		100	1	10	1	5	2	18	2
						6	1	20	1

5. Compute the mean and standard deviation of each of the following ungrouped series of X items, employing for the latter measure the equation:

$$\sigma^2 = \frac{\Sigma x^2}{N} = \frac{\Sigma X^2}{N} - M^2$$

Check by the formulas:

$$\Sigma x^2 = \Sigma X^2 - M\Sigma X; \quad \sigma = \sqrt{\Sigma x^2/N}$$

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)	(q)	(r)	(s)	(t)	(u)	(v)
3	1	4	1	1	2	2	3	1	2	1	1	2	1	1	2	1	1	2	3	1	2
5	7	8	11	3	5	12	9	7	7	7	3	6	4	5	3	4	4	2	4	4	4
5	9	8	17	4	9	12	11	9	7	10	7	15	6	5	5	5	6	6	4	10	7
11	15	20	19	5	11	12	12	15	17	13	19	24	6	5	6	8	6	8	6	11	11
				7	13	12	15	18	17	19	20	28	6	7	6	8	9	9	7	11	11
													7	7	8	10	10	9	12	11	13

6. Compute the mean and average deviation of the ungrouped series of X items in Exercise 5, by the direct formulas

$$M = \Sigma X \div N \quad \text{and} \quad AD = \Sigma |x| \div N$$

and check AD by the formula

$$AD = \frac{2(N_s M - \Sigma X_s)}{N}$$

where N_s is the number of items smaller than the mean, and ΣX_s is the sum of these items.

ANSWERS TO EXERCISES

1.

	M	Md	Mo	AD	Coef. AD	σ	Coef. σ	QD	Coef. QD
(a)	3.85	3.833	3.83	0.98	25.5%	1.195	0.310	0.925	24.2%
(b)	8.30	8.333	8.33	1.96	23.6	2.390	0.288	1.85	22.2
(c)	2.92	2.900	2.90	0.78	26.7	1.037	0.356	0.667	23.5
(d)	6.17	6.200	6.20	1.56	25.2	2.075	0.336	1.33	21.1
(e)	6.00	6.000	6.00	1.50	25.0	2.000	0.333	1.500	25.0

2.

	M	Md	Mo	AD	σ	Coef. σ	Q_1	Q_3	QD	Coef. QD
(a)	21.00	20.00	17.5	8.00	9.165	43.6%	13.75	28.33	7.29	34.6%
(b)	21.60	21.00	20.0	9.12	10.800	50.0	13.50	30.00	8.25	37.9
(c)	5.67	5.60	5.5	1.44	1.699	30.0	4.40	7.00	1.30	22.8
(d)	5.33	5.40	5.5	1.44	1.699	31.9	4.00	6.60	1.30	24.5
(e)	10.75	10.67	10.5	1.59	1.854	17.2	9.33	12.20	1.43	13.3
(f)	10.00	9.429	8.4	3.80	4.472	44.7	6.57	13.20	3.31	33.5
(g)	26.00	25.00	23.0	9.00	10.677	41.1	17.86	33.33	7.74	30.2
(h)	10.60	10.00	8.67	4.20	4.944	46.6	6.66	14.00	3.67	35.5
(i)	14.20	14.00	13.33	3.86	4.812	33.9	10.40	18.00	3.80	26.8

3.

	M	Md	Mo	AD	σ	Q_1	Q_3	QD	Coef. QD
(a)	6.9	6.600	5.67	2.49	2.96	4.556	9.000	2.222	32.8%
(b)	8.6	8.286	8.00	2.56	3.12	6.143	10.750	2.304	27.3
(c)	8.6	8.222	7.33	2.42	2.94	6.364	10.667	2.152	25.3
(d)	7.6	7.333	6.33	2.69	3.24	5.143	10.000	2.429	32.1
(e)	6.0	5.870	5.68	1.12	1.456	4.900	6.944	1.022	17.3
(f)	12.32	11.960	11.22	2.56	3.196	9.917	14.412	2.248	18.5

4.

	M	Md	Mo	GM	HM	σ
(a)	48.00	46.00	43.33	45.174	42.105	16.000
(b)	60.00	60.00	60.00	56.158	51.613	20.000
(c)	5.00	4.71	4.20	4.476	3.954	2.236
(d)	3.25	3.10	2.83	2.973	2.674	1.299
(e)	14.20	14.00	13.67	13.965	13.733	2.600

5. Means and standard deviations:

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
6	8	10	12	4	8	10	10	10	10	10
3	5	6	7	2	4	4	4	6	6	6

(l)	(m)	(n)	(o)	(p)	(q)	(r)	(s)	(t)	(u)	(v)
10	15	5	5	5	6	6	6	6	8	8
8	10	2	2	2	3	3	3	3	4	4

6. Average deviations:

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
2.5	4	5	6	1.6	3.6	3.2	3.2	5.2	5.6	4.8

(l)	(m)	(n)	(o)	(p)	(q)	(r)	(s)	(t)	(u)	(v)
7.6	8.8	1.6	1.3	1.6	2.6	2.3	2.6	2.3	3.6	3.6

B. PROBLEMS

7. The following table summarizes the number of years of experience, the average weekly sales (in units of $1,000), and the scores made in a psychological test by a group of 10 salesmen in a certain wholesale business. Find the average and standard deviation of each column of figures.

Employee	Weekly sales in $1,000 X_0	Psychological test score X_1	Experience in years X_2
A	5	4	5
B	4	5	2
C	5	6	4
D	6	4	9
E	9	5	8
F	10	6	4
G	9	6	10
H	12	7	11
I	11	9	10
J	9	8	7

8. In a certain town a sample of 176 residences was selected at random, the valuation of each residence was noted, and the distribution was tabulated as shown below. Plot the distribution in the usual rectangular form, and discover the average, median, and modal valuation. Measure the dispersion by means of the average, standard, and quartile deviation.

Value in thousands of dollars	Percentage of sample	Value in thousands of dollars	Percentage of sample
0– 1	1.7	11–12	0.0
1– 2	3.4	12–13	2.3
2– 3	4.5	13–14	0.6
3– 4	14.8	14–15	0.0
4– 5	11.4	15–16	0.6
5– 6	20.5	16–17	1.1
6– 7	15.3	17–18	0.0
7– 8	5.6	18–19	0.0
8– 9	4.5	19–20	0.0
9–10	4.0	20–21	0.6
10–11	9.1		100.0

9. An enumeration was made of the age of the men in a group of camps of the Civilian Conservation Corps in 1934, as follows.   Measure the distribution by the median, and calculate the quartile deviation.

Age	Number	Age	Number
18–19	1863	22–23	688
19–20	2096	23–24	452
20–21	1325	24–25	395
21–22	858	25 and over	246

10. In a certain mid-western town, a tabulation of the age of residences (306 homes) was made, as shown herewith.   Plot the distribution in the usual rectangular frequency form, and also plot a frequency polygon on the logarithms of the class marks.   Calculate the mean, median, and modal age, and measure the dispersion by the quartile deviation.   Also plot the cumulative percentages.

Age (in years)	Percentage of sample	Age (in years)	Percentage of sample
0– 5	3.6	55– 60	0.7
5–10	12.7	60– 65	2.3
10–15	18.6	65– 70	0.3
15–20	14.7	70– 75	0.3
20–25	11.1	75– 80	0.7
25–30	6.9	80– 85	1.3
30–35	9.5	85– 90	0.0
35–40	5.9	90– 95	0.0
40–45	6.2	95–100	0.0
45–50	1.6	˙100–105	0.3
50–55	3.3		
			100.0

11. A government bureau buying foodstuffs for relief purchased quantities of butter at several prices, as indicated.   Find the average price.

Pounds	Prices (in cents)
(w)	(m)
41,004	25
53,096	24
29,607	28
67,687	23
20,795	21

.**12.** The following data gathered by the Bureau of Labor Statistics indicate the size of families found in a number of New Hampshire towns (*Monthly Labor Review*, March, 1936, page 557).   Measure the distribution by calculating the average size of family for the area, the median, and the average deviation.

Town	Number of families (f)	Members per family (m)
Manchester	147	3.83
Nashua	100	4.02
Concord	99	3.42
Berlin	100	4.08
Portsmouth	95	3.81
Keene	97	3.41
Dover	98	3.60
Laconia	100	3.46
Claremont	100	3.51
Littleton	99	3.47
Conway	99	3.77

13. The following data (*Statistical Abstract of the United States, 1935*, page 39) describe the age distribution of the population of the United States according to two recent censuses.

(a) Prepare a chart picturing these distributions and comparing the proportions in each group at each period.

(b) Prepare another chart of the ogive type, showing the cumulative percentages up to and including the highest in the two periods ($L_2 = 5, 10, 15$, etc.).

(c) Calculate the median and quartiles of each distribution.

Age group	Per cent of total population	
	1900	1930
Under 5 years	12.1	9.3
5– 9	11.7	10.3
10–14	10.6	9.8
15–19	9.9	9.4
20–24	9.7	8.9
25–29	8.6	8.0
30–34	7.3	7.4
35–39	6.5	7.5
40–44	5.6	6.5
45–49	4.5	5.7
50–54	3.9	4.9
55–59	2.9	3.8
60–64	2.4	3.1
65–69	1.7	2.3
70–74	1.2	1.6
75–79	0.7	0.9
80–84	0.3	0.4
85 and over	0.2	0.2
Unknown	0.3	0.1
	100.1	100.1

14. The following table summarizes the simple and cumulative distribution of individual income-tax returns for 1929, by net income classes, as reported by the Federal Government. Plot the percentage distribution allowing for variable class intervals, and measure it by calculating the median and the quartile deviation.

Net income classes (Thousands of dollars)	Net income		Cumulative distribution under class above	
	Amount	Per cent	Amount	Per cent
Under 1 (estimated)	$ 73,742,132	0.30	$ 73,742,132	0.30
1 under 2 (estimated)	1,499,907,745	6.05	1,573,649,877	6.35
2 under 3 (estimated)	1,958,594,897	7.90	3,532,244,774	14.25
3 under 5 (estimated)	4,572,596,263	18.44	8,104,841,037	32.69
5 under 10	4,481,575,786	18.07	12,586,416,823	50.76
10 under 25	4,025,233,375	16.23	16,611,650,198	66.99
25 under 50	2,174,458,126	8.77	18,786,108,324	75.76
50 under 100	1,646,476,000	6.64	20,432,584,324	82.40
100 under 150	770,536,078	3.11	21,203,120,402	85.51
150 under 300	1,087,409,737	4.38	22,290,530,139	89.89
300 under 500	628,228,889	2.53	22,918,759,028	92.42
500 under 1000	669,877,752	2.70	23,588,636,780	95.12
1000 and over	1,212,098,784	4.88	24,800,735,564	100.00
Total	$24,800,735,564	100.00		

15. Lubin, in his study of the extent and duration of technological unemployment, discovered the following facts with respect to workers who found jobs within the time covered by the study:

Length of time unemployed	Number of workers
Under 1 month	47
1– 2 months	66
2– 3 months	66
3– 4 months	60
4– 5 months	43
5– 6 months	30
6– 7 months	28
7– 8 months	23
8– 9 months	18
9–10 months	10
10–11 months	7
11–12 months	3
	401

(a) Prepare charts showing: (1) the comparative success in finding positions in each month, and (2) cumulative percentages employed from month to month.

(b) Calculate: (1) the average period of unemployment of these workers, (2) the median period of unemployment, and (3) the modal period of unemployment.

(c) The standard and quartile deviation.

(d) Plot the cumulative percentage of workers against L_2 on probability paper.   Is the distribution logarithmic?

16. In an attempt to discover the possible relationship of the age of workers to their prospects for re-employment, a recent study analyzed a group of 727 unemployed workers and discovered the following age distribution (data adapted from Lubin's study):

Age group (L_1)–(L_2)	Number (f)
16–20	71
21–25	119
26–30	123
31–35	155
36–40	114
41–45	76
46–50	28
51–55	28
56–60	13

(a) What is the average age of these workers?

(b) What is the median age?

(c) What is the modal age?

(d) Locate percentiles, P_{10}, P_{30}, P_{75}.

(e) Calculate the average deviation, standard deviation, and quartile deviation.

(f) Prepare a cumulative chart or ogive portraying this distribution.

(g) From the census obtain age distributions of industrial workers and compare the distribution with that given above.   What conclusions may be drawn from this comparison?

CHAPTER VII

THE VARIABILITY OF SAMPLES

Largely on account of the expense of collecting complete data upon the facts of demand, markets, price changes, and other economic conditions, business is commonly obliged to place great reliance upon various sampling methods. In only a minority of the cases in which statistical analysis is applied can complete data be secured and subjected to investigation. Moreover, many types of statistical analysis which do not involve sampling may be closely related to it. Thus a study of demand for certain products within a small village may later be made the basis for judgments about similar characteristics of other small communities.

In any such case, analysis of limited data is made the basis for conclusions or estimates with reference to a much larger "universe" or "population." The data analyzed are regarded as a sample of the larger field. This is the process of *statistical inference*. It involves a process of inferring from the known features of available data the similar characteristics of a larger field or population. This "population" is not necessarily a group of people. A statistical population, universe, or field may be made up of persons, prices, production records, interest rates, wages, costs of living, test scores, intelligence quotients, index numbers of these conditions, or any other, similar quantitative measures. Moreover, it is not merely these actual measures, extensive as they may be, for a statistical universe is usually regarded as infinite rather than finite. It represents all the given measures there would be if the causes which occasion these measures operated freely. In other words, the statistical universe is hypothetical rather than real.

The essential problem in sampling is to secure from the

147

sample approximate measures of the various characteristics of the universe under consideration. Somewhat more realistically, the real problem is to know how reliable the measures secured by sampling may be, i.e., what are the probable limits within which these estimates secured by sampling may reasonably be expected to vary.

Obviously, means by which statistical inferences are drawn represent one of the most important considerations in a study of statistical principles and methods. Indeed, the use of sampling is so essential in economics and business, where complete data are seldom available, that sampling problems are widely regarded as among those of greatest importance in modern business statistics. It is necessary, therefore, to consider the most important characteristics of samples, their appropriate uses, and their major limitations, for there are many opportunities for the effective use of sampling techniques, but there are also serious hazards in the improper use of these devices.

In most instances, sampling is used to provide estimates or approximations of the mean, standard deviation, and other simple measures of the universes or fields from which the samples are drawn. This chapter considers the problems of sampling involved in securing approximations of these simple measures.

Random samples.—It is evident that a sample will be misleading if it is drawn in any way that reflects or introduces a bias. For example, suppose that effort is made in a city of 200,000 to estimate the numbers of adults in the population that are "prospects" for new cars. Inquiries are made of 100 persons stopping at automobile salesrooms, and it is found that 40 per cent are prospective customers. Obviously, this percentage should not be taken as representative of the city as a whole, since the fact that these persons were found in automobile showrooms indicates their probable bias.

In such a case, the information sought can be obtained only if the sample fairly represents or typifies the universe under consideration. Several methods are used to insure that this essential condition will be met. Possibly the most common seeks to insure that a *random* sample will be taken. In such usage, the sample is drawn in a manner insuring that *each and every*

member of the population has the same chance of being selected. In other words, effort is made to prevent any selective influence or bias in drawing the sample. The sample so drawn is regarded as a *random* sample.

Sometimes, however, it appears inconvenient or perhaps impossible to assume the random nature of the sample. Resort is then had to one of several other methods of sampling. The most common of these is usually referred to as *stratified* sampling. Other, less common procedures generally involve the use of various controls, which are conditions known to be closely associated with the characteristics being measured.[1]

Random sampling has been described, in preceding paragraphs, largely in negative terms. It may be said, however, that the fact that the sample is simply a bias-free selection from the universe at large should not be taken to mean that the process is a simple one. On the contrary, selection of a truly random sample requires careful consideration. Steps must be taken to insure that every member of the population actually has an equal chance of being drawn. This requirement means that sampling must avoid all influences tending to make certain members more readily available. The problems thus encountered may be more apparent if attention is directed to the procedure involved in securing a stratified sample.

Stratified samples.—It was observed in the discussion of averages that, when a statistical field is broken up into varying constituent groups, a mean may be misleading unless the data are suitably distributed among these groups. This point may be simply illustrated by the problem of sampling an electorate before an election. Until recently, such a sample was frequently secured by selecting names at random from a telephone directory, a magazine subscription list, or similar sources. Thus, for instance, every tenth name in the telephone directory might be taken. If, let us say, 10,000 names were thus sampled, and their political preferences recorded, the error would be so small

---

[1] For more detailed discussion of these specialized sampling techniques, see A. L. Bowley, "The Application of Sampling to Economic and Sociological Problems," *Journal of the American Statistical Association,* 31 (193–196), September, 1936, pp. 474–480.

that unless there was a very close vote the state of public opinion could be accurately estimated from the sample.

In recent years, however, such methods of sampling have proved entirely inadequate. The difficulty has been that political issues have in a measure "stratified" the electorate, so that sampling from a telephone directory tended to introduce an element of prejudice and to overemphasize the opinions of the well-to-do. On the other hand, in such a sample, the opinions of a large number of relief workers, laborers, or similar groups not having telephones might be almost wholly unrepresented. Similar situations might arise in many problems of sampling in economics, as, for example, in market research.

In such cases, if bias is to be avoided, it becomes necessary to resort to what is known as a stratified sample. In the stratified sampling process, the field or universe to be investigated is first of all classified according to significant differences that feature the members. For example, to take a simplified case, it might be assumed that in a given rural area (including towns up to 2,500 population) the following distribution of occupational classes is represented:

Retired landowners	10%
Farm owner-operator	30%
Village merchants	10%
Tenant farmers	20%
Laborers	30%

If these subgroups are distinctively different in terms of the characteristic under consideration, if, for instance, their consumption habits vary consistently and significantly, then it is essential that any sample taken shall give each of these occupational groups the position of prominence its size deserves. If the problem is one of forecasting election results, then it is essential: (1) that each of these subgroups be sampled; (2) that the total sample be divided so that each class is adequately and proportionately represented, and (3) that results be weighted in the over-all sample according to these proportions. Thus, by taking a sample of each group, the percentages favoring each given candidate are enumerated. A weighted average of these

percentages is then taken, the weights being the relative number of voters in each group.   The calculation takes the following form, in which expressions of opinion favoring candidate A are tabulated:

Group	Per Cent for A	Weight	Product
Retired landowners....	75	10	750
Farm owner-operator..	70	30	2,100
Tenant farmers.......	50	20	1,000
Merchants...........	45	10	450
Laborers.............	40	30	1,200
		100	)5,500

Weighted average   55%

In a broader study, various areas may be similarly classified (rural, small town, city, etc.), and the results may again be averaged with weights representing the numbers of voters in each territorial classification.   Such a process is obviously more likely to result in a close estimate than any simple selection of names from readily available lists in which various biases appear.[1]

Many other illustrations of the need for stratified sampling might be cited.   For example, the average assessment ratio (ratio of assessed to market value) of properties in a given city may be misleading if derived from an ordinary sample, because this ratio tends to "regress" or decline in the more valuable properties.   Similarly, comparisons of death rates for specified areas may be highly misleading unless they are adjusted for the age groups these localities represent.   Wealth also tends to stratify by age groups.   In fact, so common is "stratification" or "regression" that the investigator must always be on his guard against it.

In sampling with respect to public opinion, market attitudes, and similar subjects, many other difficulties arise in addition to the classification of the groups involved.   Sometimes the

---

[1] If the sample taken from each group is a fixed percentage, say 10 per cent, so that the samples are proportionate to the groups, no weighting is required.   The group samples are added to obtain the total sample.

required information cannot be obtained from simple alternative choice questions.   Somewhat elaborate questionnaires may be required.   It is obvious that such questionnaires, like those discussed in Chapter II, must be most carefully worded so as not to be ambiguous, misleading, or leading.

The normal curve of distribution.—It has been said that the most practical problem of sampling is to determine the reliability of samples.   To what extent can the sampling measures be relied upon?   Study of the variability and dependability of samples implies some sort of norm or standard by which these characteristics may be judged.   One such standard is provided in the so-called normal curve of distribution, and that standard is so widely used that it deserves extensive consideration.[1]

Tabulation of grouped data in preceding chapters has incidentally called attention to the types of distributions commonly encountered in statistical work.   An inspection of the histograms used to represent these distributions will show that many of them tend to approximate what is called a bell-shaped curve. That is, the frequencies tend to increase up to the modal frequency and to diminish in a somewhat regular succession from there on.   In some cases the figure thus suggested is symmetrical, that is, the modal frequency is at the center of the distribution, and the two sides, positive and negative, are very much alike.   In others, the plotted distribution appears to be

---

[1] The normal curve may be regarded, in terms of mathematical theory, as the limiting case of what are known as binomial distributions.   These distributions are discrete, the X scale having distinct class marks, as 0, 1, 2, etc.   The distributions may be derived by expanding algebraic binomials such as $(p + q)^n$, where $p + q = 1$.   If $p = \frac{1}{2}, q = \frac{1}{2}$, and $n = 4$, the distribution is (to avoid fractions, f is multiplied by 16)

$$m: \quad 0, \quad 1, \quad 2, \quad 3, \quad 4$$
$$2^4 f: \quad 1, \quad 4, \quad 6, \quad 4, \quad 1$$

The distribution is skewed if p and q are unequal, as $p = 0.6$ and $q = 0.4$.   In that event the distribution for $n = 4$ is

$$m: \quad 0, \quad 1, \quad 2, \quad 3, \quad 4$$
$$5^4 f: \quad 81, \quad 216, \quad 216, \quad 96, \quad 16$$

Other distributions, with varying degrees of skewness, may similarly be developed to fit special conditions.   These distributions are useful, but technical, and their treatment is deferred to a later chapter.

"skewed," that is, the frequencies are more numerous and extend farther from the mode on one side than on the other. Sometimes the skewness is positive, extending farther from the mode toward the right of the chart than toward the left. But occasionally negative skewness (greater extension to the left) is encountered. (For measures of skewness, see page 135.)

The tendency of many distributions to approximate a somewhat regular bell-shaped curve is the basis of much statistical theory. Such distributions are representative of many features of current biological, social, economic, and other natural phenomena. Thus, heights of adults in most populations show this general distributive form, as do many other characteristics of human beings. The distribution is sometimes referred to as a *chance distribution*, or its graphic representation may be called a *chance curve* or the *normal curve of error*, because it represents results that would follow from the operation of certain common accidental or chance probabilities. It has been found that the typical distribution may be represented by a mathematical formula, and this formula may also be taken as an expression of the laws of chance.[1] It is not convenient at this point to consider the detailed mathematical characteristics of this type of distribution, but the theoretical normal distribution to which actual distributions tend to conform may be briefly described.

The theoretical normal curve as mathematically calculated may be regarded as a composite picture of numerous actual distributions, in which the distributions have been reduced to a

[1] If the magnitude scale (x) is in standard deviation units, the normal curve is

$$Y = e^{\frac{-x^2}{2}}$$

where e is the growth constant $2.71828+$, x is a deviation from the mean, and the modal ordinate is 1. On semi-log paper the curve becomes a parabola, and on double-log paper the logarithmic normal curve becomes a parabola. In more general terms the equation of the curve of unit area is

$$Y = (2\pi)^{-0.5} e^{\frac{-x^2}{2\sigma^2}}$$

Thus if σ is taken as unity, Y at $x = 1$ becomes

$$Y = (1 \div \sqrt{2 \times 3.14159}) \, (1 \div \sqrt{2.71828})$$

$$= 0.39894 \times 0.60653 = 0.2420$$

which is the ordinate at $x/\sigma = 1$ in the table of the normal curve of unit area.

standardized scale, with skewness eliminated. It is, therefore, descriptive of a general form that real distributions commonly approximate. As thus considered, it has many useful applications in problems of variability and probability.

The normal curve, illustrated in Fig. 7·1 differs from commonly used distributions of grouped data in that the horizontal unit is most conveniently expressed in terms of the standard deviation,[1] and the classes are so reduced in width that the class
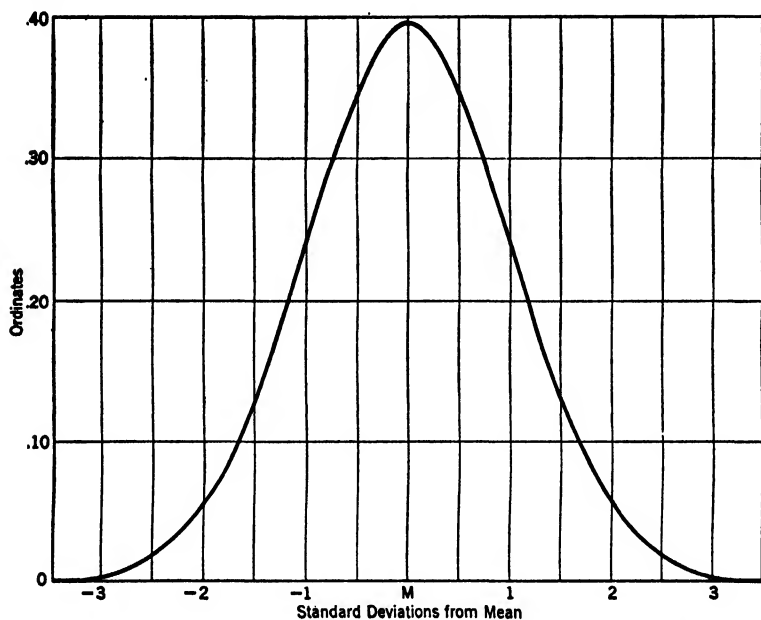
Fig. 7·1.—The Normal Curve, Showing the Mean and Standard Deviations.

interval is of negligible size. Further, the curve is not generally described as a succession of frequencies, but rather as a succession of areas. These areas are thought of as representing a series of very narrow classes, each of which has as its class interval a small fractional part of a standard deviation and as its vertical measure the height of the ordinate at the theoretical class mark. Frequencies are regarded as areas, i.e., width times height.

---

[1] Some statisticians, particularly those working in the field of education, designate the x scale in σ units (that is, x/σ) as z. In this usage $10z + 50 = T$, or $T = (10x/\sigma) + 50$.

Mathematicians have computed tables of the ordinates and areas of the normal curve such as that illustrated in Table 7·1

TABLE 7·1

THE NORMAL CURVE OF DISTRIBUTION

σ units $\pm x$	Height, z	Area from 0 to $+x$	Area from $-x$ to $+x$	σ units $\pm x$	Height, z	Area from 0 to $+x$	Area from $-x$ to $+x$
0.0	0.3989	0.0000	0.0000	2.1	0.0440	0.4821	0.9643
0.1	0.3970	0.0398	0.0797	2.2	0.0355	0.4861	0.9722
0.2	0.3910	0.0793	0.1585	2.3	0.0283	0.4893	0.9786
0.3	0.3814	0.1179	0.2358	2.4	0.0224	0.4918	0.9836
0.4	0.3683	0.1554	0.3108	2.5	0.0175	0.4938	0.9876
0.5	0.3521	0.1915	0.3829	2.6	0.0136	0.4953	0.9907
0.6	0.3332	0.2257	0.4515	2.7	0.0104	0.4965	0.9931
0.7	0.3122	0.2580	0.5161	2.8	0.0079	0.4974	0.9949
0.8	0.2897	0.2881	0.5763	2.9	0.0060	0.4981	0.9963
0.9	0.2661	0.3159	0.6319	3.0	0.0044	0.4986	0.9973
1.0	0.2420	0.3413	0.6827	3.1	0.0033	0.4990	0.9981
1.1	0.2178	0.3643	0.7287	3.2	0.0024	0.4993	0.9986
1.2	0.1942	0.3849	0.7699	3.3	0.0017	0.4995	0.9990
1.3	0.1714	0.4032	0.8064	3.4	0.0012	0.4997	0.9993
1.4	0.1497	0.4192	0.8385	3.5	0.0009	0.4998	0.9995
1.5	0.1295	0.4332	0.8664	3.6	0.0006	0.4998	0.9997
1.6	0.1109	0.4452	0.8904	3.7	0.0004	0.4999	0.9998
1.7	0.0940	0.4554	0.9109	3.8	0.0003	0.4999	0.9999
1.8	0.0790	0.4641	0.9281	3.9	0.0002	0.5000	0.9999
1.9	0.0656	0.4713	0.9426	4.0	0.0001	0.5000	0.9999
2.0	0.0540	0.4772	0.9545				

$$\sigma = 1.0000 \qquad \pm QD \text{ includes 50 per cent of area}$$
$$AD = 0.7979\sigma \qquad \pm 1.96\sigma \text{ includes 95 per cent of area}$$
$$QD = 0.6745\sigma \qquad \pm 2.576\sigma \text{ includes 99 per cent of area}$$

(for a more complete table see Appendix, page 582). It will be seen that the total area of the curve is always unity. In such terms, since the product of height and width is area, the height of the ordinates (z), even at the mode, is necessarily fractional (0.3989). (The ordinates are also frequently expressed as percentages of the maximum ordinate, Y_0, taken as unity.) It will

be noted that the six standard deviation units in the horizontal measurement include nearly all (99.7 per cent) of the area. Thus the height multiplied by width, in standard deviation units, equals area, which is expressed as a percentage of the total area, which is 1, or unity. This area, of course, can be multiplied by a constant to suit the convenience of any special case. Since the curve is symmetrical, both sides being the same, tables generally refer only to one half the curve, and that is true of the tables just mentioned.

These known characteristics of the normal curve make many estimates, forecasts, and predictions possible. Thus, when the mean and standard deviation (the essential measures) of a given field are known, it is possible to estimate limits within which measures of a sample selected from the field are likely to fall.

If the distribution from which a single random item is drawn is normal, it can be stated, by reference to Table 7·1, that there are 34 chances out of 50, or approximately 68 in 100, that it will be within the limits of one standard deviation below and above the mean (at 1.0 in first column, read 0.3413 in third column and 0.6827 in fourth column). Hence, in general it may be stated that the chances are that a random item 68 times out of 100 will be within one standard deviation of the mean, either above or below it.

In the same way it may be determined from the table that the chances are 0.4773 out of 0.5000 or approximately 95 out of 100 that a random item will fall within the limits of two standard deviations above and below the mean (at 2.0 in first column read 0.9545 in fourth column). By careful interpolation of the curve it may be shown that the exact 95 per cent probability just mentioned (fourth column) is really at 1.96σ instead of 2σ (first column). Further, it may be shown that there are 997 chances in 1,000 (0.9973 in fourth column), or practical certainty, that the random item will fall within the limits of three standard deviations above and below the mean. By interpolation, again, this probability may be reduced to express exactly the 99 per cent probability, namely, that there are 99 chances in 100 that the random item will fall within the limits of 2.576σ about the mean. The 95 and 99 per cent probabilities at 1.96

and 2.576, respectively, are the ones most commonly referred to in tables dealing with normal probabilities. They are usually described, not in terms of the probability of the item's being included within the given limits, but of its being found outside these limits, at a greater deviation from the mean. Hence, they are called the 5 per cent and 1 per cent levels of probability, respectively, implying that there is a 5 per cent chance that a random item will fall outside the $\pm 1.96\sigma$ limits and a 1 per cent chance that it will fall outside the $\pm 2.576\sigma$ limits. These limits may also be described as the *fiducial* or *confidence* limits, implying faith in the probability (95 or 99 per cent) that a random item will fall within the stated limits.

The probabilities involved in connection with various x/σ measures, as the 5 per cent and 1 per cent levels, may be illustrated by reference to a specific investigation which involved a study of incomes. On the basis of a complete census rather than a sample, it was reported that, in round numbers, the mean of family incomes for a given region was $1,500, and the standard deviation was $280. It appeared reasonable to assume in this case that the distribution of incomes was practically normal. What probable statements might be made regarding the range of incomes?

In the first place, it could reasonably be concluded that 68 per cent of the incomes were within one standard deviation above and below the mean, that is, between the limits

$$1\sigma \text{ limits} = \$1,500 \pm \$280$$

$$1\sigma \text{ range: from } \$1,220 \text{ to } \$1,780$$

In the same way 2σ and 3σ limits could be determined. But perhaps it would be more useful to compute the 5 per cent and 1 per cent levels, that is, the limits including 95 per cent and 99 per cent of the incomes, respectively, as follows:

$$5\% \text{ level, or } 1.96\sigma \text{ limits } = \$1,500 \pm (1.96 \times \$280)$$
$$95\% \text{ range: from } \$951 \text{ to } \$2,049$$

$$1\% \text{ level, or } 2.576\sigma \text{ limits } = \$1,500 \pm (2.576 \times \$280)$$
$$99\% \text{ range: from } \$779 \text{ to } \$2,221$$

Stated in terms of probabilities these limits mean that there is a 5 per cent chance that a random income from this field will fall outside the approximate limits of \$951 and \$2,049 and a 1 per cent chance that such an individual case will fall outside the limits \$779 and \$2,221.

In practice limits such as have just been set would often be inexact because many actual distributions exhibit varying degrees of skewness. Income distributions, for instance, are usually positively skewed. In such cases, the limits as computed would be too low. It is possible to make allowance for skewness.[1] As will be explained in a later section, however, these computed probabilities are generally regarded as approximate measures of reliability, and their strict application should be confined to such measures as tend to distribute themselves normally. This is, for instance, generally true of the means of successive large samples drawn from a given distribution.

Variability among samples.—As has already been observed, statistical analysis makes wide use of samples. Often it is convenient to depend entirely on a sample for evidence as to the characteristics of a given statistical field or universe from which the sample is drawn. There is, however, always a danger of making erroneous inferences if the sample is small, so that *statistical induction*—a term denoting this process of making inferences—must

---

[1] An adjustment for skewness suitable for many positively skewed distributions having a lower limit at zero makes use of the geometric mean and the logarithmic standard deviation. That is, M and σ are calculated from the logs of m instead of from m (see pages 132 and 133). $M \pm \sigma$, $M \pm 2\sigma$, etc., may then be written as a simple progression. Their antilogs form a geometric progression and represent the required limits corrected for skewness. The measure of central tendency is thus the geometric mean, and the standard deviation intervals expand above and contract below the mean in geometric progression. Thus the lower limits of $M \pm 3\sigma$, based on logs, cannot extend below zero, while the upper limit may have a much wider range, as determined by the degree of skewness. For instance, the arithmetic (A), logarithmic (L), and geometric (G) progressions to $\pm 3\sigma$ for the distribution on p. 119 may be summarized as follows:

	$M - 3\sigma$	$M - 2\sigma$	$M - \sigma$	M	$M + \sigma$	$M + 2\sigma$	$M + 3\sigma$
A	9.09	10.97	12.84	14.72	16.60	18.47	20.35
L	0.9967	1.0526	1.1084	1.1643	1.2202	1.2761	1.3320
G	9.92	11.29	12.84	14.60	16.60	18.88	21.48

(Logarithmic $\sigma = 0.055877$; geometric σ ratio $= 1.1373$.)

be undertaken with caution. Extensive analysis has, for this reason, been directed toward the problems of sampling and the characteristics of samples, particularly as to the degree to which their means, standard deviations, and other measures tend to represent and to vary from those of the so-called parent population or statistical universe from which these samples are taken.

By way of emphasis, it may be repeated that the terms "population" and "universe," as used in connection with sampling theory, refer primarily to a statistical field from which a random sample is drawn, with the understanding that the "parent" field is very large. For example, an intensive study might be made of a small group of migratory unskilled workers in the United States, on the assumption that it will furnish information pertinent to all such workers. Yet the concept of "population" or "universe" is even more abstract than this, since it ignores the limitations of the actual statistical field. In reality, conclusions based on a random sample drawn from a normally distributed statistical population represent generalizations affirmed to be true of the same class of units wherever they might be found.

In actual statistical analysis, as has already been noted, the limitations of expense and time commonly make it impossible to obtain more than one moderately sized sample of the data under investigation. In order to illustrate the peculiar characteristics of samples, however, it is desirable here that not one but a succession of samples be drawn, so that their tendency to vary from one another and from the parent population may be observed.

A brief illustrative study of this sort appears in Example 7·1. The procedure involved in this simplified illustration may be briefly described as follows: First, a very limited parent population or statistical field was created. It consisted of the series of numbers or magnitudes listed at the top of the example. Theoretically, of course, this statistical field should have comprised thousands of items of magnitudes such as to represent a normal distribution. However, the limited "parent population" used in this case effectively serves the purpose at hand, which is merely that of illustration. Individual numbers com-

prising this statistical field were written upon metal-rimmed tags and shuffled in a container. A sample of 5 items each was then obtained as follows: One tag was drawn, its number recorded, and the tag again returned to the container. The tags were again shuffled and a second item was drawn, the number recorded, and the tag again returned. In the same way, 3 other items were drawn to complete the sample of 5 random items. It is necessary that each tag be returned to the container before a second is drawn so that each magnitude will have an equal chance of being selected. If the universe were of vast extent, this procedure would be unnecessary.

<div align="center">EXAMPLE 7·1</div>

<div align="center">VARIABILITY AMONG SAMPLES</div>

Data: The limited statistical population: 3; 7; 9; 11; 13; 14; 15; 16; 17; 18; 19; 19; 21; 21; 22; 22; 23; 23; 24; 25; 25; 26; 27; 27; 28; 28; 29; 29; 31; 31; 32; 33; 34; 35; 36; 37; 39; 41; 43; 47. $M = 25$; $\sigma^2 = 97.5$; $\sigma = 9.8742$. Also 10 random samples, $a \ldots j$, of 5 items each, with their M's and σ's.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
	29	15	15	35	22	11	24	23	31	15
	15	3	14	33	34	7	39	29	16	14
	15	29	25	28	41	34	37	47	23	16
	21	32	47	7	27	27	21	32	28	43
	19	24	28	25	23	19	22	27	41	13
Σ	99	103	129	128	147	98	143	158	139	101
M	19.8	20.6	25.8	25.6	29.4	19.6	28.6	31.6	27.8	20.2
								(Ave. $M = 24.9$)		
σ	5.15	10.52	11.92	9.95	7.17	9.95	7.76	8.24	8.33	11.44
								(Ave. $= 9.043$)		

<div align="center">Standard deviation of sample means: 4.28</div>

<div align="center">Same, theoretical: $9.8742 \div \sqrt{5} = 4.42$</div>

After the 10 samples of 5 items each had been recorded, the mean and standard deviation of each sample were computed. The variability of the means thus obtained suggests how reliable any one sample is likely to be as a basis for determining the true mean of the statistical field from which it is drawn. What are

the chances that a single sample would have a mean close to the true mean of the field?

The answer to the question just propounded obviously depends on the extent of variability among the sample means, and this in turn depends upon the variability of the statistical field.   If the sample means are all practically alike, it is highly probable that any sample will give a close approximation to the mean of the parent population or "universe" (M_u).   But if they vary widely, not much reliance can be placed on the mean of a single random sample as representative of the true mean.

Variability among the means may be measured by finding their standard deviation by the usual method.   This standard deviation of the 10 means is found to be 4.28, which, of course, must not be confused with the standard deviation of any one sample, or of the whole statistical field ($\sigma_u = 9.874$).   It appears, therefore, that, if the mean of the statistical field is to be judged from the first sample drawn, the mean of this sample (19.8) might be considered a variable with a standard deviation of 4.28.   It is clearly a somewhat uncertain quantity.   If, however, the standard deviation of many sample means had been very small, the mean of the first sample would be relatively dependable.

Inference of mean and standard deviation.—A consideration of the procedure by which the standard deviation of the sample means (σ_M) was obtained will readily suggest that this standard deviation has a definite relation to the standard deviation of the statistical field as well as to the size of the sample.   The larger the variability of the field, the more the sample means might be expected to vary; but the larger the size of each sample, the less the means are likely to vary.   The relation between the standard deviation of a very large number of sample means (σ_M) and the standard deviation of the universe from which these samples are taken at random (σ_u), however, has been subjected to extensive study and may be expressed in the formula

$$\sigma_M = \frac{\sigma_u}{\sqrt{N}}$$

where N is the number of items in each sample.

In most statistical practice, however, the standard deviation of the universe is unknown and must most commonly be estimated from a sample. In this estimation, the standard deviation of the sample is recognized as probably smaller than that of the universe, because in random drawings items close to the mode are more likely to be drawn than those in the tails of the distribution.[1] It has been found that the standard deviation of the sample may generally be adjusted so that it more nearly approximates the standard deviation cf the universe as

$$\sigma_u = \sigma_{adj} = \sigma_s \sqrt{\frac{N}{(N-1)}} = \sqrt{\frac{\Sigma x^2}{N-1}}$$

The statistic[2] thus provided represents what is described as the "best estimate" of the standard deviation of the universe that may be derived from this sample. In the case of sample a in Example 7·1 (page 160), for instance, where the standard deviation of sample (a) is 5.1536, the best estimate of the standard deviation of the universe from this sample is

$$\sigma_u = 5.1536 \sqrt{\frac{5}{(5-1)}} = 5.1536\sqrt{\tfrac{5}{4}} = 5.762$$

It will be clear from this illustration that the standard deviation thus estimated is not necessarily the true σ_u. In the long run, however, it tends to approximate the true measure more closely than does the simpler standard deviation cf a sample. The latter is said to have a "downward bias," which is illustrated in Example 7·1, where the standard deviations of the various samples average 9.043, while the standard deviation of the universe is 9.874.

[1] That σ_s tends to be smaller than σ_u may be more definitely shown. Assume that, from a normal universe expressed as $x = X - M$, a sample of N items (Σx_s) is drawn. Then $\sigma_s^2 = (\Sigma x_s^2/N) - M_s^2$. But $\Sigma x_s/N = M_s$ is an unbiased but variable estimate of the mean of x, or zero, and the correction term, M_s^2, therefore indicates a downward bias, which becomes smaller as N is increased.

[2] The term *statistic* is often used to describe a measure derived from a sample, such as a sample mean or standard deviation. The corresponding measure derived from the entire universe is called a *parameter*. The statistic, either directly as M, or indirectly as $\sigma_s\sqrt{N/(N-1)}$, may be considered an estimate of a parameter but not a perfect measure of it.

Adjustment of the standard deviation derived from the sample is, of course, reflected in the measure of the standard error of the mean. As has been noted,

$$\sigma_M = \frac{\sigma_u}{\sqrt{N}}$$

When the adjusted standard deviation is substituted to provide a superior estimate of the standard deviation of the universe, this formula may be restated as

$$\sigma_{M_{adj}} = \frac{\sigma\sqrt{N/(N-1)}}{\sqrt{N}} = \sqrt{\frac{\Sigma x^2}{N(N-1)}}$$

The last expression is known as *Bessel's formula*. It defines σ_M when the only evidence obtainable with respect to σ_u is that secured from a sample.[1]

Degrees of freedom.—Brief further attention may well be given to the nature of the adjustment by which the standard deviation of the sample is made the best estimate of the standard deviation of the universe and σ_M is similarly corrected. That adjustment illustrates an important feature of modern statistical theory, in which, where dependence is placed on sampling, N is corrected to become what is described as "degrees of freedom." In the case at hand, for instance, there are said to be $N - 1$ degrees of freedom, and the adjustment may be rationalized by comparing the process involved with that of computing a sample mean. In the case of the mean, the first sample item

---

[1] Standard errors of sampling are computed for unlimited *universes* or *populations*, but they may be adapted approximately to limited populations by the use of the following correction factor (K) applied to the squared standard error as ordinarily computed:

$$K = \frac{N_p - N_s}{N_p}$$

Thus, supposing that a sample of $N_s = 100$ is drawn from a limited population of $N_p = 1000$, and it is estimated from the sample that M is 50, and σ_M^2 as ordinarily computed is 27.778. Then the corrected σ_M^2 is

$$\sigma_{M_c}^2 = 27.778 \times \frac{1000 - 100}{1000} = 25.00$$

and

$$\sigma_{M_c} = \sqrt{25.00} = 5.0$$

drawn constitutes a preliminary crude estimate, and each additional sample item theoretically improves that estimate. Hence the divisor for the sum of all items is simply N. In the case of the standard deviation, however, no conclusions are possible from the first sample item drawn. A second sample item is required as a basis for the first crude estimate of the standard deviation. Hence, in measuring deviations, the count begins with the second sample item rather than the first. Similarly, therefore, the divisor for the sum of squared deviations becomes $N - 1$ rather than simply N. This is a somewhat over-simplified explanation of the degrees-of-freedom concept. The real proof of its propriety, however, requires complex mathematical demonstration which is not appropriate here.

Reliability of the sample mean.—The next question that may logically be raised in connection with the sampling process described in preceding pages concerns the definition of limits within which the true mean of the universe may be expected to lie on the basis of information gained from the sample. In other words, when the measures of the sample are known, when the statistics σ and σ_M have been found, what rules if any govern the relationship between the mean of the sample and that of the universe? It will be clear that this question is essentially concerned with the reliability of the sample mean as an estimate of the mean of the universe.[1]

---

[1] The standard sampling error generally varies as the square root of N. When this is the case, N therefore varies as the square of the standard error. Hence in such a case it is possible to estimate from a given sample the size of the sample required to reduce a standard error to a given value.

Formula: Given a sample of N_1 items, and a standard error of the mean (E_1) computed from it (where E varies as $\sqrt{N}$), to find the sample size (N_2) required to reduce E_1 to a specified E_2:

$$N_2 = \frac{E_1^2}{E_2^2} \times N_1$$

Illustration: A sample of 20 items has a mean of 13, and a σ_M^2 of 3.61053. A mean practically accurate to 10 per cent is required; that is, assuming 99 per cent probability, it is required that $2.576\sigma_M = M/10$, or $\sigma_M^2 = M^2/25.76^2 = 0.25468$.

Hence $N_2 = \dfrac{3.61053}{0.25468} \times 20 = 284$.

The accuracy of the estimate is limited by the representativeness of the sample.

In part, this question has been answered in the adjustment of the standard deviation secured from the sample and the similar correction in the statistic σ_M. To illustrate, reference may again be made to the data of Example 7·1, page 160, where the adjusted standard deviation of the first sample is found to be 5.762, and the adjusted σ_M may be found as:

$$\sigma_M = \frac{\sigma_u}{\sqrt{N}} = \frac{5.762}{\sqrt{5}} = 2.58$$

or

$$\sigma_M = \sqrt{\frac{\Sigma x^2}{N(N-1)}} = \sqrt{\frac{132.8}{5 \times 4}} = 2.58$$

The problem of interpreting this adjusted σ_M must next be considered. What does it mean? How may it be used to appraise the reliability of the sample mean?

If the true mean of the universe were known and if this σ_M were the actual parameter of the universe, then the nature of the parent distribution would be clear. These two measures would define a distribution of means that would appear from numerous samples. In that case, the usual rules of probability would prevail, so that some 68 per cent of the sample means might be expected to fall within $1\sigma_M$ of the true mean, 95 per cent within $1.96\sigma_M$, and 99 per cent within $2.576\sigma_M$.

Actually, however, both the mean and the standard error of the mean must commonly be estimated from a sample. As the data of Example 7·1 clearly indicate, the sample mean, especially if N is small, may be far from the true mean, so that the usual probabilities with respect to σ_M cannot be applied when reliance is placed on small samples. Rather, some adjustment which takes account of the errors introduced by the combination of estimates, which recognizes and compensates for the variability of small samples, must be used for this purpose. Most convenient, in this connection, is the reference to fiducial or confidence limits based, not on the assumptions of a normal distribution and the use of large samples, but on a modification of the normal distribution, a modification known as the t-distribution.

Fiducial limits based on t.—As has been noted, when the measures of a universe are unknown and must be estimated from a small sample, the simple σ_M derived from the sample cannot be depended upon accurately to define the limits within which the true mean of the parent distribution lies. The distribution of sample means implied by σ_M thus derived may not have the same mean as that of the single sample, although the mean of the sample is the best estimate of the true mean. The usual rules of probability, therefore, do not apply exactly to measures derived from small samples. For instance, in such cases, the probability that the mean of an additional sample of the same size and from the same parent universe will fall within $1.96\sigma_M$ of the sample mean is not 95 per cent. Rather, the 95 per cent limits must include, in such cases, a somewhat wider range to compensate for the uncertainty concerning the estimated σ_M. Similarly, the 99 per cent limits must include somewhat more than the usual $2.576\sigma_M$. How much wider this range must be depends primarily upon the size of the sample.

The table of t, of which an abbreviated form is illustrated in Table $7 \cdot 2$, has been prepared in order to answer the question of how wide a range must be included within the usual confidence limits when those limits are based on samples of various sizes. It may be noted there, for instance, that, for samples involving only 5 items, the 95 per cent limits are $2.776\sigma_M$ instead of the usual $1.960\sigma_M$; and the 99 per cent limits are $4.604\sigma_M$ instead of the usual $2.576\sigma_M$. It will be apparent that such a small sample requires broad adjustments. For samples of 40 items each, however, these limits are notably narrowed, while samples approaching the infinite in size show the usual non-sampling probability limits, as may be noted in the table. For other values of t, see last column, F-table, page 586.

The usefulness of the t distribution may be demonstrated and the approach to sampling problems of this type may be summarized at the same time by reference to a simple illustration. It may be assumed that a firm desires a reasonably accurate estimate of the average savings of its employees. It does not, however, wish to arouse the antagonism that might appear if all workers were questioned. For that and other

reasons, it places dependence upon a sample of 10 cases. The mean of the sample is found to be \$350, while Σx^2 is 36,000, and the standard deviation is \$60. The mean of the sample is accepted as the best estimate of the true mean, but question is raised as to its reliability, that is as to the probable limits within which the true mean lies.

TABLE 7·2

ABBREVIATED TABLE OF LIMITS FOR t *

Probabilities		Number of Items (N) and Degrees of Freedom (n)			
Inside	Outside	$N=5$ $n=4$	$N=10$ $n=9$	$N=20$ $n=19$	$N=\infty$ $n=\infty$
0.50	0.50	$0.741\sigma_M$	$0.703\sigma_M$	$0.688\sigma_M$	$0.674\sigma_M$
0.60	0.40	0.941	0.883	0.861	0.842
0.70	0.30	1.190	1.100	1.066	1.036
0.80	0.20	1.533	1.383	1.328	1.282
0.90	0.10	2.132	1.833	1.729	1.645
0.95	0.05	2.776	2.262	2.093	1.960
0.98	0.02	3.747	2.821	2.539	2.326
0.99	0.01	4.604	3.250	2.861	2.576
0.999	0.001	8.610	4.781	3.883	3.291

* The sampling distribution of t is symmetrical, but it departs somewhat from the normal in the case of small samples. The relative height of the curve, with $N - 1 = n$ degrees of freedom, may be defined as

$$Y_t = \left(1 + \frac{t^2}{n}\right)^{-\frac{N}{2}}$$

and the area is

$$A_t = \frac{(n\pi)^{\frac{1}{2}}[(n-2)/2]!}{[(n-1)/2]!} = \frac{(n\pi)^{\frac{1}{2}}G(n/2)}{G(N/2)}$$

where ! is the factorial sign attached to a given factor, and G is the gamma function. The standard deviation of Y_t is

$$\sigma_t = \sqrt{n/(n-2)}$$

It will be clear that, as N increases, σ_t approaches unity. The distribution then approaches the normal form. In the table, limits for $N = \infty$ are simply x/σ for the normal curve.

While large-sample theory, which assumes the adequacy of M and σ_M as estimated from a sample, has been widely utilized in approaching such problems, greater dependability might well

be placed upon a procedure that refers to the t distribution.   In that procedure, the standard error of the mean would first be estimated in the usual manner as

$$\sigma_M = \frac{\sigma}{\sqrt{N-1}} = \frac{60}{\sqrt{10-1}} = \frac{60}{3} = 20$$

Reference would then be made to the table of t and to the column "$n = 9$," that is, $N - 1 = 9$.   There, it will be noted that, for a sample of the size here taken, the 95 per cent probability requires $2.262\sigma_M$, while the 99 per cent probability requires $3.250\sigma_M$.   Accordingly, the limits within which the true mean might be expected to lie may be defined as follows:

95% limits: $M \pm 2.262\sigma_M = 350 \pm 2.262 \times 20 = 305$ to 395
99% limits: $M \pm 3.250\sigma_M = 350 \pm 3.250 \times 20 = 285$ to 415

This type of procedure is definitely more conservative than that based on the usual rules of large-sample theory.

Confidence ratio.—It is often convenient to describe the range within which the true mean probably lies by expressing half this range as a ratio to the sample mean.   Thus, assuming the data just given for 10 items ($n = 9$), and a 95 per cent probability, the following ratio may be written:

$$t_{95}\frac{\sigma_M}{M} = 2.262 \times \frac{20}{350} = 0.13 = 13\%$$

In a normal distribution this is equivalent to saying that the true mean probably lies within 13 per cent above or below the sample mean or, in other words, that the sample mean probably has an error not larger than 13 per cent.   Whether this error is too large in a given study, of course, is a matter of judgment, just as when the confidence limits are stated.   One advantage of stating the limits in the ratio form is that in cases of moderate skewness the distortion of the limits due to the skewness is avoided and at the same time an approximate indication of the accuracy of the mean is provided.

Many other related sampling variabilities or "errors" will be described in later chapters.[1]

Standard error of the difference between means.—Attention may now be directed briefly to a further application of the standard error of the mean. In it, this measure is utilized to discover whether two or more fields from which samples are available are significantly different. The setting for the problem may be briefly described as follows. Samples have been taken from each of the fields, and their means and standard deviations have been noted. The question is raised whether they represent the same parent population or whether, on the other hand, one of them has been drawn from a distinctly different universe.

To illustrate, it may be assumed that a large manufacturing company compares two sources of electric-light bulbs to discover whether there is a significant difference in the usable lifetime of their respective products. Records with respect to a random sample of 50 bulbs from one company are compared with similar records for a random sample of 50 bulbs from the second. Means, standard deviations, and standard errors of means for each of the samples are computed and may be summarized as follows:

$$\text{Firm A:} \quad M = 600 \text{ hours}; \quad \sigma_A = 63; \quad \sigma_{M_A} = \frac{63}{\sqrt{49}} = 9$$

$$\text{Firm B:} \quad M = 700 \text{ hours}; \quad \sigma_B = 56; \quad \sigma_{M_B} = \frac{56}{\sqrt{49}} = 8$$

On the basis of the simplest comparison, the products of Firm B seem more durable, since their average life is 700 hours whereas that of bulbs furnished by Firm A is 600 hours. The difference

---

[1] It may be noted that the standard deviations of the samples in Example 7·1, page 160, vary considerably. The degree of variability among such standard deviations may be estimated by the formula:

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}}$$

where N is the number of items in the sample. It may also be noted that the degree of sampling variability, that is, the standard error of sampling, of the average deviation is $0.603\sigma_M$. Of the median it is $1.253\sigma_M$, and of the quartiles it is $1.363\sigma_M$.

between the means is obviously 100 hours. That there is, however, a wide range of variability among the items in each sample is indicated by the size of the two standard deviations. The essential question may, therefore, be restated in the following form: is the variability among items within each sample so great that the difference between the two averages shows no certain or significant difference between the two parent universes from which the samples have been taken?

Where the samples are of the same size, and where there is no correlation between the two series of sample items (as there might well be were the comparison related to the performance of an identical sample of workmen before and after some change in productive technique), this question may be answered by reference to a measure that combines the standard errors of the means of the two samples. These two measures, σ_{M_1} and σ_{M_2}, are first found in the usual manner. Then the *standard error of the difference between two sample means*, σ_D, is calculated as

$$\sigma_D = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2}$$

It estimates the standard deviation of innumerable D's (i.e., $M_2 - M_1$) obtained from paired sample means of N items each. With respect to the data described in preceding paragraphs, this measure appears to be

$$\sigma_D = \sqrt{9^2 + 8^2} = \sqrt{81 + 64} = \sqrt{145} = 12.04$$

This measure of reliability is applied in a manner similar to that described in connection with σ_M. In practice, unless samples are very small, it is frequently assumed that, if the difference between two means is less than three times σ_D, the difference between parent universes is not conclusively established.

In the example just cited, the actual difference between the means, 100, is over 8 times σ_D. Such a large difference is unlikely to occur by mere chance. It may therefore be con-

cluded that the product of Firm B is definitely superior to that of Firm A.[1]

The assumptions underlying this procedure and the conclusion just described deserve brief attention.   The reasoning makes use of what is called a *null hypothesis*, which is so designated because it involves temporary reference to assumptions directly contrary to the point the procedure seeks to demonstrate, in this case the distinctive character of the samples.   In other words, it says in effect: We are seeking evidence that the two samples represent different parent populations; well, let us start by assuming that they have been drawn from the same population.   We can safely conclude on the basis of large-sample theory that, if large numbers of paired random samples of N items each are drawn from a single universe and their means are compared, the differences between these paired means $(M_2 - M_1)$ will tend to form a normal distribution with an average of 0.   The standard deviation of this distribution is σ_D. Under such circumstances, it would be expected that practically all (99 per cent) of the actual differences found in this manner would fall within the range of $\pm 3\sigma_D$.   If, therefore, the actual difference between two means exceeds $3\sigma_D$, then it may reasonably be concluded that the source of the two samples thus compared is not the same population.   As has been indicated, a more exact measure of the significance of a given difference is available in the statistic t.

This type of problem is also effectively approached by the method of variance analysis as described in Chapter XIX.   It receives further consideration in that connection.[2]

---

[1] For samples that are unequal in size the test is somewhat less valid.   A more exact test can be made by computing

$$t = \frac{|M_2 - M_1|}{\sigma_D} = \frac{700 - 600}{12.04} = 8.31$$

and comparing it with 5 and 1 per cent levels of t at $N - 2$ degrees of freedom, where N is the number of items in the two combined samples (see Table of F, last column, page 586).   The nearest tabulated values are 1.984 and 2.626, respectively.   Since the t value found is several times as large as these limits, the previous conclusion is confirmed.

[2] If in the illustration here cited the two samples (Y_1 and Y_2) had varied in size (N_1 items in first sample and N_2 in second, where $N_1 + N_2 = N$), the same method

READINGS

(See also special and general references, pages 591 and 597.)

BOWLEY, A. L., "The Application of Sampling to Economic and Sociological Problems," *Journal of the American Statistical Association*, 31 (195), September, 1936, pp. 474–480.

COATS, R. H., "Enumeration and Sampling," *Journal of the American Statistical Association*, 26 (175), September, 1931, pp. 270–284.

GALLUP, G., and RAE, S. F., *The Pulse of Democracy*, New York, Simon and Schuster, 1940, 335 pp.

ROBERTSON, W. L., "Quality Control by Sampling," *Factory and Industrial Management*, 76 (3), September, 1928, pp. 503–505; and (4), October, 1928, pp. 724–726.

STARCH, DANIEL, "Factors in the Reliability of Samples," *Journal of the American Statistical Association*, Supplement, 27 (177A) March, 1932; pp. 190–201.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Compute the standard error of the mean (the predicted variability of successive sample means) for each of the following samples of five items each, drawn from various sources:

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)
1	2	2	3	1	2	3	6	2	2	5	4	4	4	6	51
3	5	12	9	7	7	9	8	10	6	15	8	6	16	22	53
4	9	12	11	9	7	12	12	14	15	20	12	10	20	24	54
5	11	12	12	15	17	15	24	18	24	25	16	18	28	30	55
7	13	12	15	18	17	21	25	26	28	35	20	22	32	38	57

2. Using the data of Example 1, and assuming normal universes, compute the fiducial or confidence (95 per cent) range of the true mean, that is, the

---

might, under certain conditions, be utilized. But when the usual assumptions obtain, the standard error of the difference of the means is more precisely determined thus:

$$\sigma_D^2 = \frac{\Sigma y_1^2 + \Sigma y_2^2}{N - 2} \times \frac{N}{N_1 N_2}$$

which reduces to $\sigma_{M_1}^2 + \sigma_{M_2}^2$ if $N_1 = N_2$.

Where the two samples are of the same size but each Y_1 is paired with a Y_2 (so-called correlated data), as when they represent scores for the same workmen tested at different times, σ_D may most easily be found as the standard error of the mean of the "gains" (G), where $G = Y_2 - Y_1$. (Cf. Chapter XIX (page 471) and "Seasonality in Strikes," by Dale Yoder, *Journal of the American Statistical Association*, 33, December, 1938, pp. 687–693).

In this connection, also, it may be noted that the statistic F utilized in variance analysis is directly related to the t mentioned in this chapter, in that t is the square root of the corresponding F.

limits between which 95 per cent of the sample means fall, as estimated from the given sample (95 per cent range $= M \pm t_{95}\sigma_M$; or when $N = 5$: $M \pm 2.776\sigma_M$). Also compute the 95 percentage range $(t_{95}\sigma_M/M)$.

3. Assuming that Y_1 and Y_2, in each example below, represent comparable samples, determine whether there is a significant difference between their means by finding the value of $t = |M_2 - M_1| \div \sigma_D$.

(a)		(b)		(c)		(d)		(e)		(f)		(g)	
Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
3	8	1	12	3	23	1	7	1	12	2	6	4	4
5	12	7	14	9	25	3	11	3	15	4	9	6	16
6	16	9	18	12	29	7	20	4	19	5	13	10	20
7	20	10	26	15	41	19	29	5	21	6	15	18	28
9	24	13	30	21	42	20	33	7	23	8	17	22	32

(h)		(i)		(j)		(k)		(l)		(m)		(n)	
Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
1	9	1	5	3	17	7	8	6	17	5	14	6	5
2	9	4	7	7	21	7	10	8	31	8	16	14	32
4	14	6	10	10	33	12	22	20	50	9	19	26	41
5	15	6	14	12	35	13	26	24	64	12	23	42	68
5	17	6	14	13	36	15	26	30	65	12	23	42	68
7	20	7	16	15	38	18	28	32	73	14	25	50	86

Answers to Exercises

1.

(a) 1	(e) 3	(i) 4	(m) 3.464
(b) 2	(f) 3	(j) 5	(n) 4.899
(c) 2	(g) 3	(k) 5	(o) 5.292
(d) 2	(h) 4	(l) 2.828	(p) 1

2.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
L_1	1.224	2.448	4.448	4.448	1.672	1.672	3.672	3.896
L_2	6.776	13.552	15.552	15.552	18.328	18.328	20.328	26.104
%	69.4	69.4	55.5	55.5	83.3	83.3	69.4	74.0

	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)
L_1	2.896	1.120	6.120	4.149	2.384	6.400	9.309	51.224
L_2	25.104	28.880	33.880	19.851	21.616	33.600	38.691	56.776
%	79.3	92.5	69.4	65.4	80.1	68.0	61.2	5.1

3. Computed t (cf. table at $n = 10 - 2$ or $n = 12 - 2$):

(a) 3.333	(e) 6.261	(i) 3.000	(m) 4.472
(b) 3.000	(f) 3.131	(j) 5.000	(n) 1.425
(c) 4.000	(g) 1.333	(k) 2.000	
(d) 1.562	(h) 5.000	(l) 3.000	

B. Problems

4. A certain firm employing a sales organization in each of two states wished to determine whether there was a significant difference between the efficiency of the two organizations. Random samples of 10 items each were available, as listed below. Determine whether the difference between the means is significant (data simplified for purposes of illustration).

First organization		Second organization	
Salesman	Dollar sales (000)	Salesman	Dollar sales (000)
A	1	K	7
B	4	L	8
C	6	M	5
D	5	N	4
E	2	O	3
F	8	P	7
G	10	Q	12
H	1	R	10
I	2	S	5
J	1	T	9

5. In an effort to reduce the number of defective washing machines produced in its plant, a manufacturing concern instituted a system of inspection covering each of the major operations in production. The comparative performance before and after the institution of the inspection system, as measured by the number of units rejected by distributors, is shown in the following table, where rejects per week are tabulated for two periods of 12 weeks each, the one preceding and the other following the provision of inspectors. General production conditions throughout the whole 24 weeks were strictly comparable except for the change in the inspection system, as noted.

UNITS REJECTED BY DISTRIBUTORS, PER WEEK

Week	Inspection not provided	Week	Inspection provided
1	44	13	39
2	56	14	41
3	73	15	20
4	27	16	24
5	39	17	35
6	72	18	41
7	50	19	33
8	60	20	38
9	66	21	29
10	65	22	34
11	51	23	56
12	69	24	18

Evaluate the statistical significance of the provision of inspectors. Is the evidence conclusive? What are the assumptions upon which this analysis is based?

6. The S. H. Company, securities brokers, maintains a company school for additions to its sales force. Only a limited number of its recruits are able to attend, however, and the company is interested to discover whether there is an appreciable difference in the performance of those who attend and those who are unable to do so. The following tabulation compares average monthly sales of new salesmen in their first two years of full-time employment with the company. The two groups are not appreciably different, except with respect to the training, as noted.

AVERAGE MONTHLY SALES
(In thousands of dollars)

Graduates of the training program		Salesmen untrained in the company school	
Salesman	Sales	Salesman	Sales
A	4	A	6
B	5	B	6
C	3	C	3
D	4	D	4
E	6	E	2
F	8	F	2
G	2	G	3
H	5	H	4
I	9	I	2
J	4	J	3
		K	9
		L	4

Show that there is or is not a significant difference between the two groups that might justify the company in continuing the program.

7. The personnel office of the T Company has been requested to furnish a statement of the average amount paid annually by its 2,800 non-clerical employees as life-insurance premiums. Data available in the personnel files, and regarded as representing a small random sample, indicate the following payments for 20 employees:

Annual premium	Number of employees	Annual premium	Number of employees
$6.00	1	$12.00	6
8.00	2	15.00	2
9.00	4	25.00	1
10.00	3	45.00	1

Using these data, the personnel office seeks to estimate how large a sample must probably be taken in order to secure, within fiducial limits of $3\sigma_M$, a sample average within 10 per cent of the theoretical true average.

Note: From the available sample σ_u may be estimated as $\sqrt{\Sigma d^2 \div (N-1)}$

CHAPTER VIII

INDEX NUMBERS

Any extensive examination of literature dealing with economics and business, or for that matter, with many other aspects of modern society, will reveal the widespread use of index numbers. Changes in prices, production, wages, costs of living, employment, growth of population, and in numerous other fields are frequently represented by index numbers. Thus an index number is a means by which data gathered in various times and places may be readily compared. The process of comparison may be facilitated by expressing the variables as percentages of some common base, either a given date or a given period or place.

An index of prices.—The index of wholesale prices in the United States as computed by the Bureau of Labor Statistics is summarized in Table 8·1, and the annual figures, together with estimates for years prior to 1890, appear in Fig. 8·1. This index measures price changes from year to year and from month to month in a representative list of important wholesale commodities. It is also reported currently by weeks. The figures are quoted as percentages of the average prices prevailing during the year 1926, which is called the base year. Hence, the figure 64.8, which is the index of wholesale prices in 1932, means that in that year, on the average, important commodities in wholesale markets sold at 64.8 per cent of their selling price in 1926. In 1940, they sold at an average of about 78.5 per cent of their cost in the base year.

The wholesale prices of all important commodities for which

177

TABLE 8·1

WHOLESALE PRICES IN THE UNITED STATES, 1890–1939 (1926 = 100)

Source: Bureau of Labor Statistics.

Annual averages

1890 = 56.2	1900 = 56.1	1910 = 70.4	1920 = 154.4	1930 = 86.4
1891 = 55.8	1901 = 55.3	1911 = 64.9	1921 = 97.6	1931 = 73.0
1892 = 52.2	1902 = 58.9	1912 = 69.1	1922 = 96.7	1932 = 64.8
1893 = 53.4	1903 = 59.6	1913 = 69.8	1923 = 100.6	1933 = 65.9
1894 = 47.9	1904 = 59.7	1914 = 68.1	1924 = 98.1	1934 = 74.9
1895 = 48.8	1905 = 60.1	1915 = 69.5	1925 = 103.5	1935 = 80.0
1896 = 46.5	1906 = 61.8	1916 = 85.5	1926 = 100.0	1936 = 80.8
1897 = 46.6	1907 = 65.2	1917 = 117.5	1927 = 95.4	1937 = 86.3
1898 = 48.5	1908 = 62.9	1918 = 131.3	1928 = 96.7	1938 = 78.6
1899 = 52.2	1909 = 67.6	1919 = 138.6	1929 = 95.3	1939 = 77.2

Monthly averages

	1926	1927	1928	1929	1930	1931	1932
January	103.2	96.5	96.4	95.9	92.5	78.2	67.3
February	102.0	95.8	95.8	95.4	91.4	76.8	66.3
March	100.6	94.7	95.5	96.1	90.2	76.0	66.0
April	100.3	94.1	96.6	95.5	90.0	74.8	65.5
May	100.5	94.2	97.5	94.7	88.8	73.2	64.4
June	100.4	94.1	96.7	95.2	86.8	72.1	63.9
July	99.5	94.3	97.4	96.5	84.4	72.0	64.5
August	99.1	95.2	97.6	96.3	84.3	72.1	65.2
September	99.7	96.3	98.6	96.1	84.4	71.2	65.3
October	99.4	96.6	96.7	95.1	83.0	70.3	64.4
November	98.4	96.3	95.8	93.5	81.3	70.2	63.9
December	97.9	96.4	95.8	93.3	79.6	68.6	62.6
Year	100.0	95.4	96.7	95.3	86.4	73.0	64.8

	1933	1934	1935	1936	1937	1938	1939
January	61.0	72.2	78.8	80.6	85.9	80.9	76.9
February	59.8	73.6	79.5	80.6	86.3	79.8	76.9
March	60.2	73.7	79.4	79.6	87.8	79.7	76.7
April	60.4	73.3	80.1	79.7	88.0	78.7	76.2
May	62.7	73.7	80.2	78.6	87.4	78.1	76.2
June	65.0	74.6	79.8	79.2	87.2	78.3	75.6
July	68.9	74.8	79.4	80.5	87.9	78.8	75.4
August	69.5	76.4	80.5	81.6	87.5	78.1	75.0
September	70.8	77.6	80.7	81.6	87.4	78.3	79.1
October	71.2	76.5	80.5	81.5	85.4	77.6	79.4
November	71.1	76.5	80.6	82.4	83.3	77.5	79.2
December	70.8	76.9	80.9	84.2	81.7	77.0	79.2
Year	65.9	74.9	80.0	80.8	86.3	78.6	77.2

reasonably accurate quotations are readily obtainable are included in this particular series, which now includes between 800 and 900 price series. Compiling data for such an index from the principal markets of the country requires a well-organized reporting system and a carefully arranged procedure which defines the qualities of the goods included in the index and determines methods of selection for the figures to be recorded. The index number is "broken down" into various groups of com-
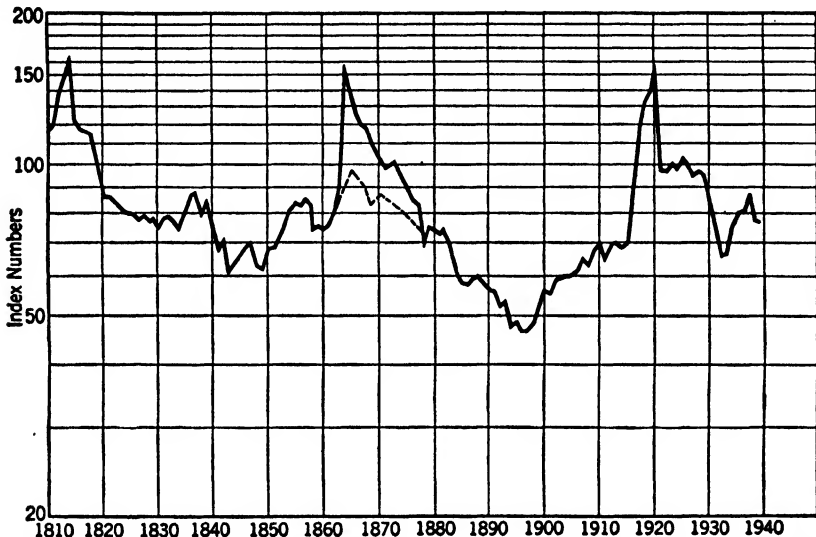
Fig. 8·1.—Index Numbers of Wholesale Prices in the United States, 1810–1939. (1926 = 100.) Source: United States Bureau of Labor Statistics.

modities, and indexes for these various subdivisions are also published. Figure 8·2 illustrates this procedure as it applies to the cost-of-living index prepared by the Bureau of Labor Statistics. The figure compares indexes for each of the six component series and the general index.

CALCULATION OF INDEX NUMBERS

Selecting a base.—Although the term *base* may have more than one meaning, it ordinarily implies the period chosen to represent 100 per cent. No absolute rules can be stated for the selection of a base for a series of index numbers or relatives, ·
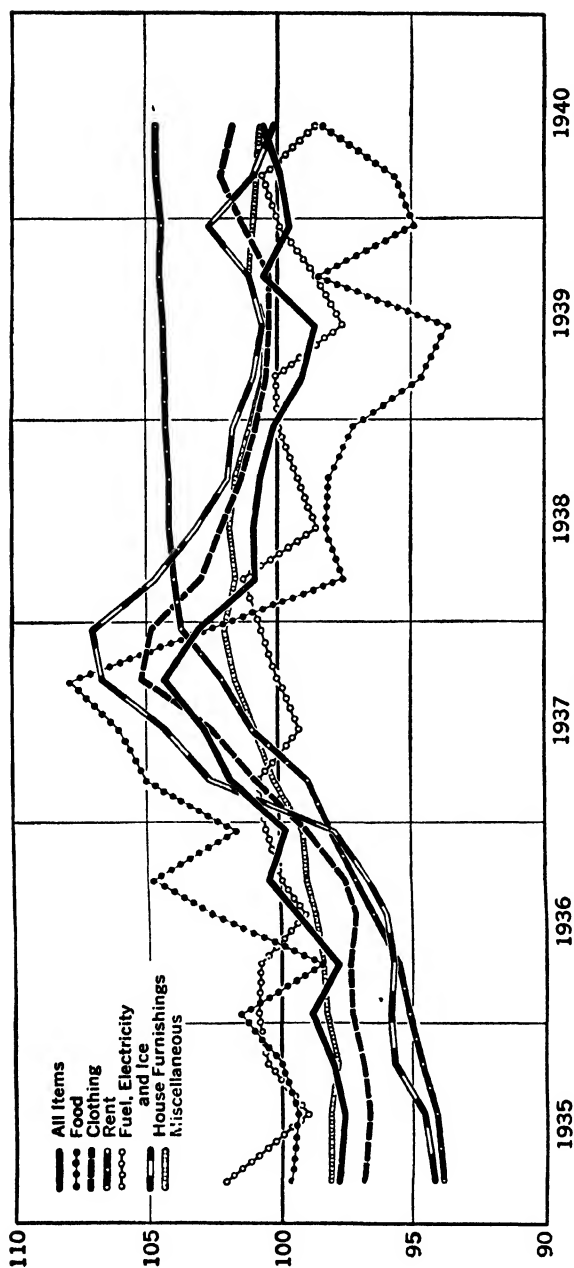
FIG. 8·2.—Index Numbers of the Cost of Goods Purchased by Wage-Earners and Lower-Salaried Workers in Large Cities, 1935–1940, by Classes of Goods. (Average 1935–39 = 100.) Source: United States Bureau of Labor Statistics.

but several considerations deserve mention. In the first place, it should be noted that the base may be a single year or it may be a composite of several years. Many business series are, for instance, based on the average of the years 1923 through 1925, or that of 1935 through 1939. Usually, the base is selected to represent what is believed to be a fairly typical or normal period in which few unusual factors influenced the data under consideration. In other cases, however, where data have not been long available or where no satisfactory judgment as to what is typical is possible, the first year's figures may be taken for basing purposes. Again, where data are to be used for some specific comparison, it is desirable to select a particular period conformable with this objective. Thus, if the relatives are to show changes since pre-war or pre-depression periods, 1913 or 1928 might reasonably be selected as a base. It is well to keep in mind, however, that the more distant the base, the less dependable are comparisons with that base.

Finally, it should be understood that, while the selection of an appropriate base is frequently important, it is usually possible to adjust bases from one period to another without serious difficulty, as will be indicated in a subsequent section of this chapter.

A simple index.—In order to center attention upon the basic principle involved in the calculation of index numbers, reference will first be made to the method of securing relatives for a single series. Table 8·2 demonstrates the construction of such an index.

In column 2 of Table 8·2 are summarized the numbers of workers involved in industrial disputes in the United States for the period 1916 to 1939. The numbers are fairly large, and comparison of one year with another is made difficult by this fact. In column 3, however, the number of workers involved in each year has been stated simply as a percentage of the number in the first year, which is taken as the base for the series. This column represents a series of simple index numbers obtained by dividing each given number by the number in the base year and expressing the result as a percentage, that is, multiplying it by 100. Thus, the index is $(1,466,695 \div 1,599,917) 100 = 92$

for 1934. Comparison of the different years is clearly facilitated by reference to these relatives.

TABLE 8·2

WORKERS INVOLVED IN INDUSTRIAL DISPUTES, 1916–1939

Year	Numbers	Index numbers or relatives (1916 = 100)
1916	1,599,917	100
1917	1,227,254	77
1918	1,239,989	78
1919	4,160,238	260
1920	1,463,054	91
1921	1,099,247	69
1922	1,612,562	101
1923	756,584	47
1924	654,641	41
1925	428,416	27
1926	329,592	21
1927	349,434	22
1928	357,145	22
1929	230,463	14
1930	158,114	10
1931	279,299	17
1932	242,826	15
1933	788,995	49
1934	1,466,695	92
1935	1,117,213	70
1936	788,648	49
1937	1,860,621	116
1938	688,376 ·	43
1939	1,170,962	73

Composite index numbers.—Most index numbers in common use, however, are not simple in the sense that they involve only one series of data. For instance, reference may be made to the abbreviated index of farm prices in the middle west, presented in Example 8·1, which illustrates the calculation of a composite index. By an appropriate weighting process, an

average price is first calculated for the several commodities in 1929, and a similar average of these prices is also calculated for 1932. Relatives or index numbers are then computed, in order to measure the change of prices between the two dates, by expressing the later data as a ratio to the earlier. Cost-of-living indexes are another illustration of this type of comparison.

EXAMPLE 8·1

A PRICE INDEX, COMMON AGGREGATIVE METHOD

Data: Prices of selected agricultural products and typical pre-depression marketings in the middle west.

Com-modities	Units (thou-sands of)	Typical market-ings q_w	1929		1932	
			Prices p_0	Product p_0q_w	Prices p_1	Product p_1q_w
Cattle	cwt.	187	$10.78	$2,015.86	$4.91	$918.17
Hay	ton	6	10.66	63.96	7.60	45.60
Corn	bushel	1,411	0.77	1,086.47	0.23	324.53
Butter	pound	2,697	0.46	1,240.62	0.20	539.40
Eggs	dozen	2,078	0.26	540.28	0.11	228.58
Totals			$\Sigma p_0 q_w = 4{,}947.19$		$\Sigma p_1 q_w = 2{,}056.28$	
Index numbers			$\dfrac{\Sigma p_0 q_w}{\Sigma p_0 q_w} = 100\ (\%)$		$\dfrac{\Sigma p_1 q_w}{\Sigma p_0 q_w} = 41.6\ (\%)$	

The aggregative method.—The method illustrated in the example just referred to is merely an adaptation of a weighted average. The weights selected are the quantities of these commodities ordinarily marketed. As will be seen later, it would not do to use variable weights representing the actual marketings of each given period of time, since this would introduce into the result a factor representing quantity changes, whereas it is required that price changes alone should be measured. Hence, identical weights, representing typical marketings, are used for both periods.

From the standpoint of a weighted average, the prices represent the measure (m), and the quantities represent the frequencies or weights; that is, they represent the number of times the price would be expended under normal conditions. The tabulation, however, does not take the form of a frequency distribution, nor are class intervals involved. Nevertheless, the average may be obtained in accordance with the usual formula,

$$M = \frac{\Sigma mw}{\Sigma w} \quad \text{or} \quad \frac{\Sigma mf}{N}$$

although the division by N is omitted as unnecessary, because the average price at the two successive dates is in itself unimportant. What is important is the ratio of the average price at the later period to the average price at the initial period,[1] and this ratio is the same for the Σmw's as for the M's; that is (using subscripts 1 and 2 to indicate the first and second years, respectively),

$$\frac{\Sigma m_2 w}{\Sigma w} \div \frac{\Sigma m_1 w}{\Sigma w} = \frac{\Sigma m_2 w}{\Sigma m_1 w}$$

In Example 8·1 the total Σmw for 1929 is found to be $4,947.19, and for the later date, $2,056.28. These aggregates may themselves be regarded as index numbers, since they represent the composite change in prices. Their importance, however, lies in their ratio to each other; and they may each be multiplied or divided by the same number, since this does not change their ratio. If each is divided by the 1929 aggregate ($4,947.19), they become 100 (per cent) and 41.6 (per cent), respectively. As has been noted, the year or other period thus represented by 100 is called the base, and a year compared with it is called a *given* year. In references to such index numbers the per cent sign is generally omitted.

---

[1] The question may arise, why should not prices be weighted by the actual quantities, and the effect of a change in quantity be discounted by dividing by Σw, or N. This is the method which obviously would be used in averaging the prices of successive purchases of the same commodity. The answer is that such a method would give variable results depending upon the units employed, as pounds or tons, feet or yards, etc. Hence, it would be indeterminate. But if units could be standardized this would be the logical method. See footnote on next page.

This system of computing an index number is called the *common aggregative* method. It is one of the most frequently used methods, although it is not entirely above criticism from a strictly mathematical standpoint. Such criticism generally points to the fact that the method in effect averages numbers that do not logically constitute the basis for an average, since they represent prices of diverse commodities. The validity of the method as a most useful and workable approximation, however, has been established by comparison with more logically justifiable methods and by experience as well.[1]

An index of quantity.—It is frequently desirable to measure changes in the physical volume of production, marketing, or other business activity from time to time, irrespective of price changes in the same period. An index that permits such comparisons is called a *quantity index*. Such an index is so con-

[1] It has often been objected that the quantities held constant should not be regarded as weights. It is said that they represent diverse units (hundredweights, bushels, yards, etc.) which cannot properly be added. As in other cases, however, such a procedure is justified by the logic of the process (e.g., cross products in correlation, page 331), which reduces the items thus treated to abstract numbers. Pricing is essentially a market process which renders commensurable the physical units of diverse commodities in terms of equal values, such as a dollar's worth. Thus the data of Example 8·1 may be rewritten making quantities the number of dollar's worth marketed at standard (p_0) prices, thus:

1929			1932	
q_w	p_0	$p_0 q_w$	p_1	$p_1 q_w$
2,015.86	1	2,015.86	0.45547	918.16
63.96	1	63.96	0.71295	45.60
1,086.47	1	1,086.47	0.29870	324.53
1,240.62	1	1,240.62	0.43478	539.40
540.28	1	540.28	0.42308	228.58
4,947.19		4,947.19		2,056.28
Weighted averages:		$p_o = 1.00$		$p_1 = 0.416$
Index numbers:		100		41.6

As thus stated the weights (q_w) may properly be added, and the weighted averages of p_o and p_1, written as percentages, represent the required price index numbers. But if the original typical marketings are applied as abstract weights to the prices, the same index numbers are more conveniently obtained as ratios of the weighted totals.

structed as to take account of the number of tons, bushels, kilowatt-hours, or other units which must be averaged or aggregated to provide a composite measure. An excellent example
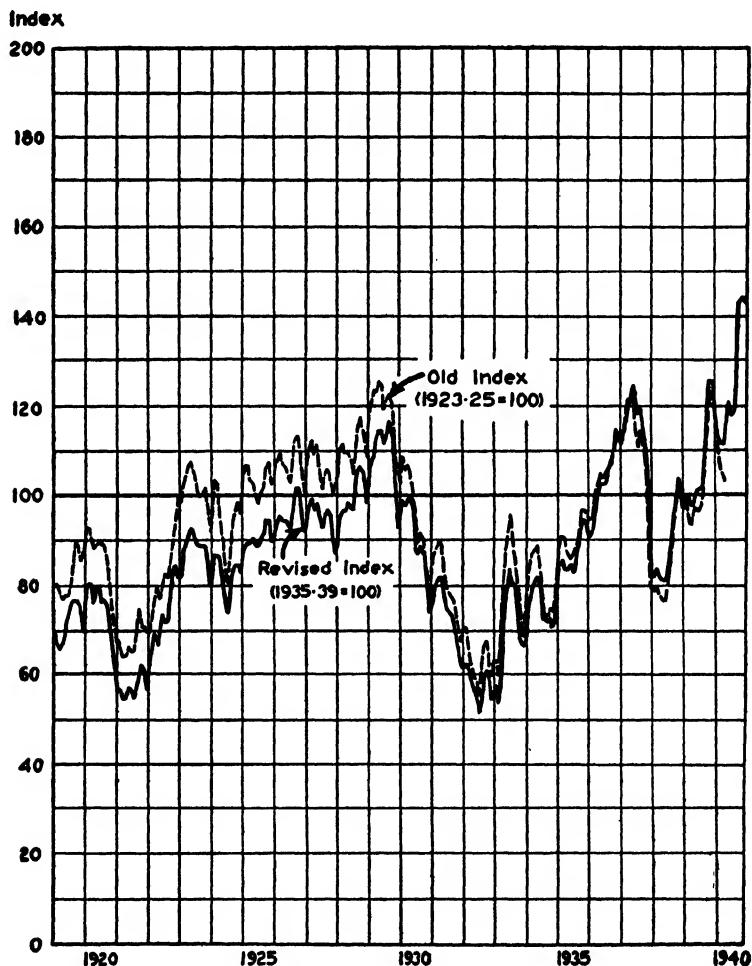
FIG. 8·3.—Federal Reserve Indexes of Industrial Production in the United States; Old Index and 1940 Revision. Source: *Federal Reserve Bulletin*, Vol. 26, No. 8, August, 1940, p. 754.

of such an index is the Index of Industrial Production compiled by the Federal Reserve Board. This index as revised in 1940 (see Table 8·3 and Fig. 8·3) brings together a wide range

TABLE 8·3
INDEXES OF INDUSTRIAL PRODUCTION, UNITED STATES, 1923-1940

(1940 revision, unadjusted for seasonal change, 1935–1939 average = 100.)
Source: *Federal Reserve Bulletin*, August, 1940, and succeeding issues.[1]

Month	1923	1924	1925	1926	1927	1928	1929	1930	1931
January........	82	83	87	91	93	91	103	96	75
February.......	85	87	89	94	97	95	108	100	79
March.........	89	87	90	96	100	97	110	98	81
April..........	91	84	90	95	97	97	113	100	82
May..........	93	81	91	95	99	99	115	99	82
June..........	92	77	89	95	97	98	115	95	78
July..........	89	74	89	93	93	97	112	88	75
August........	89	78	91	98	96	102	114	87	74
September......	89	83	92	102	97	106	117	89	73
October........	89	85	95	102	96	107	114	86	70
November......	86	85	95	98	91	104	103	80	67
December......	80	83	90	92	87	99	93	74	63
Annual index	88	82	91	96	95	99	110	91	75

Month	1932	1933	1934	1935	1936	1937	1938	1939	1940
January........	62	56	69	80	91	112	82	98	117
February.......	63	58	75	85	91	115	82	99	113
March.........	62	54	79	86	94	120	84	100	112
April..........	59	59	81	84	100	122	82	98	112
May..........	57	69	82	84	103	125	81	99	116
June..........	55	79	80	85	103	120	81	102	121
July..........	52	84	73	84	103	118	85	102	118
August........	54	81	73	87	106	120	90	103	120
September......	60	80	72	90	108	115	95	116	129
October........	62	74	73	94	111	110	99	126	134
November......	59	68	71	95	114	97	102	126	135
December......	55	67	74	94	114	86	100	124	135
Annual index	58	69	75	87	103	113	88	108	122

[1] The entire index, with comments on the revision and with detailed series unadjusted and adjusted for seasonality, is available in the *Survey of Current Business*, 20(8), August, 1940, pp. 11–18. Items for 1919 to 1922, inclusive, chained from the unrevised index are also included. For a criticism of the new index see *Cleveland Trust Company Business Bulletin*, September 15, 1940, p. 4.

of data representing the output of shops, factories, mines, and various processing units. The general index is published in the *Federal Reserve Bulletin* and in the *Survey of Current Business*, as are subsidiary indexes representing its major components. The index is broken down into two major divisions, manufactures and minerals, which are based upon 81 individual production series, of which 28 were added to the earlier (1927) index. The division described as manufactures includes indexes representing iron and steel, automobiles, food products, aircraft, furniture, rayon, and many others, while the minerals division presents indexes of coal, iron-ore, lead, petroleum, gold, silver, etc. Agriculture and building construction are not directly represented in the index, although changes in these fields appear indirectly, since agricultural fluctuations are reflected in the foods division and changes in construction show up in connection with the various materials entering into such construction.

In order that relatives involved in the index may be more strictly comparable from month to month, data are often reduced to average output per working day, and a parallel series of relatives, adjusted for seasonal variations, is always available from the same sources. The significance of this adjustment will be given special attention in a subsequent chapter.

The essential principle involved in the construction of a *quantity index*, or an *index of physical volume*, as it is frequently called, is similar to that featuring the price index already described. Production figures are recorded as if for averaging, with typical prices in the given markets as weights. These prices or weights must be held constant from one date to another, just as quantity weights are held constant in Example 8·1. As in Example 8·1, also, it is not necessary to divide by the sum of the weights, since the aggregates themselves express the required ratios. The procedure, therefore, differs from that of Example 8·1 only in that the position of prices and quantities is reversed. That is, in a price index, actual prices are multiplied by constant (typical) quantities; in a quantity index, actual quantities are multiplied by constant (typical) prices.

The method of calculating a quantity index is illustrated by reference to brief data in Example 8·2.

<div align="center">

EXAMPLE 8·2

A QUANTITY INDEX, COMMON AGGREGATIVE METHOD

</div>

Data: Approximate production and typical prices of five commodities (United States). Prices in 1926 arbitrarily accepted as typical.

Commodity	Units (millions omitted)	Typical Prices	1928		1929	
			Production	Value	Production	Value
		p_w	q_0	$p_w q_0$	q_1	$p_w q_1$
Pig iron	Long tons	$21.00	38	798.0	42	882.0
Bituminous coal	Short tons	4.30	501	2,154.3	535	2,300.5
Cement	Barrels	1.74	176	306.24	170	295.8
Corn	Bushels	0.75	2,819	2,114.25	2,535	1,901.25
Wheat	Bushels	1.45	915	1,326.75	813	1,178.85
Product totals			$\Sigma p_w q_0 = 6,699.54$		$\Sigma p_w q_1 = 6,558.40$	
Index numbers			$\dfrac{\Sigma p_w q_0}{\Sigma p_w q_0} = 100\ (\%)$		$\dfrac{\Sigma p_w q_1}{\Sigma p_w q_0} = 97.9\ (\%)$	

Longer time periods.—Methods of computing index numbers of price and quantity described in preceding paragraphs are the ones most commonly employed. As has been suggested, the definition of data and their collection, together with the choice of appropriate weights, involve the exercise of discretion and judgment, so that comparable index numbers emanating from different sources will be subject to some variation.

In Examples 8·1 and 8·2, comparison involves two years only. The same method, however, may be repeated for successive years or for convenient subdivisions, such as months or weeks. If weights are selected for months, separate weights for each month are preferable, and a similar procedure is desirable

in comparing weekly data. These variable weights are, of course, the typical figures for the designated periods. When an index is continuous for many years, the weights are frequently revised in order to keep them more closely in line with what is typical. Such revision may be made at stated intervals, as every five or ten years, or as occasion may seem to demand. Over very long periods of time, index numbers relating to such data as costs of living, if they are not so revised, may not have great significance because of changes in buying habits. For that reason, and in an effort to reduce the errors occasioned in changing weights, most business indexes are commonly limited to staple commodities, such as basic foodstuffs and metals.

An index of value.—In practice, index numbers of prices (P) and of physical quantities (Q) are computed for many markets. But so-called indexes of values produced or exchanged (V) are also sometimes calculated.

A value index in its simplest form may be illustrated by the index of sales which is sometimes computed by individual firms or corporations, or even by separate departments of such concerns.[1] The aggregates for such an index may readily be obtained from the accounting department in the form of sales totals, in dollars. Returned merchandise is usually deducted if significant in amount and of variable proportions. The aggregates thus obtained may readily be reduced to index numbers expressed as a percentage of a selected base, just as was done for price and quantity indexes.

It is important to observe that the original data represented by the sales aggregates consists of units of goods listed at the actual sales price. The problem of selecting weights does not arise. The form of calculation from original data is therefore as illustrated in Example 8·3, though in practice, as just indicated, the aggregates might be available directly.

Value indexes might also be prepared for data on business, such as bank debits for cities, states, or the nation as a whole. These figures represent bank debits to individual accounts, that

---

[1] The *Survey of Current Business* carries indexes of retail sales in small towns and rural areas, which may be cited as indexes of value. It also carries aggregates of income in dollars for the nation, which might readily be reduced to index numbers.

is, checks drawn and cashed. Since the large majority of these checks represent sales (Σpq), the aggregates may be regarded as a rough indication of business change, though they may be distorted by excessive speculative transactions, bankruptcy transfers, or other unusual items. When divided by total deposits in corresponding banks, they also throw some light on the rate of currency circulation.

EXAMPLE 8·3

A VALUE INDEX: SALES OF CLOTHING

Data: Abbreviated data of goods sold and selling prices in Department X of a certain department store for designated periods. Quantities (q) sold, selling price (p), and value (v) for each grade of goods.

Item	1938 (base)			1939								
				January			February			March		
	q	p	v	q	p	v	q	p	v	q	p	v
Suits:												
Grade A	31	42.00	1,302.00	25	40.00	1,000.00	15	38.50	577.50	27	37.50	1,012.50
Grade B	65	32.50	2,112.50	61	32.50	1,982.50	45	30.00	1,350.00	64	29.00	1,856.00
Grade C	87	24.75	2,153.25	85	23.50	1,997.50	51	22.75	1,160.25	86	20.00	1,720.00
Grade D	88	19.50	1,716.00	90	18.00	1,620.00	48	15.00	720.00	91	16.00	1,456.00
Overcoats:												
Grade A	18	45.00	810.00	15	42.25	633.75	10	38.75	387.50	20	35.00	700.00
Grade B	25	82.50	812.50	24	30.00	720.00	13	27.50	357.50	26	25.00	650.00
Grade C	35	25.00	875.00	30	22.50	675.00	18	20.00	360.00	38	18.50	703.00
Totals	Σpq = $9,781.25			Σpq = $8,628.75			Σpq = $4,912.75			Σpq = $8,097.50		
Index numbers	100.00			88.22			50.23			82.79		

"Splicing" and linking.—Within reasonable limits it is permissible to change the base of a series of index numbers by dividing each item by the one which is to be reduced to 100. This procedure is based upon a recognition of the nature of index numbers as a series of ratios, so that a common multiplier or divisor has no effect on their interrelationships.   The process is merely an extension of the principle that both terms of a fraction may be simultaneously multiplied or divided by a constant without altering the value of the fraction.   For example, reference may be made to the index numbers of wholesale prices in

the United States for the years 1926–1932 inclusive. They
may be summarized as follows:

$$1926 = 100.0$$
$$1927 = \ \ 95.4$$
$$1928 = \ \ 96.7$$
$$1929 = \ \ 95.3$$
$$1930 = \ \ 86.4$$
$$1931 = \ \ 73.0$$
$$1932 = \ \ 64.8$$

Suppose that for certain reasons it is desirable to emphasize
comparisons of all other years with 1929, i.e., to reestablish the
series upon a 1929 base.   The desired change may be made by
dividing each index by the index for 1929, namely, 95.3, in which
case the series appears as follows.   (For convenience each divi-
dend is simultaneously multiplied by 100, thus retaining the
percentage form.)

YEAR	INDEX (1926 = 100)			INDEX (1929 = 100)
1926	100.0	÷	95.3	=	104.9
1927	95.4	÷	95.3	=	100.1
1928	96.7	÷	95.3	=	101.5
1929	95.3	÷	95.3	=	100.0
1930	86.4	÷	95.3	=	90.7
1931	73.0	÷	95.3	=	76.6
1932	64.8	÷	95.3	=	68.0

Since indexes may thus be changed by means of a common
factor or divisor, it is possible to combine into a single index
partial indexes which overlap.   For example, suppose that there
has been computed in a certain city an index of the cost of living
for the years 1926 to 1930, as follows:

COST OF LIVING, CITY A, 1926–1930

1926	100
1927	98
1928	96
1929	96
1930	93

Suppose, also, that this index was discontinued and another
index purporting to cover the same field of prices was calculated
for the years 1930 to 1933, inclusive, as follows:

Cost of Living, City A, 1930–1935

1930	100
1931	90
1932	82
1933	81

These two indexes may be combined into a single index by changing either of them so that it agrees with the other in 1930. Thus, each item in the first series may be divided by 0.93 (or multiplied by $1/0.93 = 1.0753$), so bringing the series to 100 in 1930, in agreement with the second. Or, if it is desirable to preserve the original base (1926), each item in the second series might be multiplied by 0.93 so that the series, on a 1926 base, would be:

1930, 93; 1931, 84; 1932, 76; and 1933, 75

as appended to the first series. By either process, the two indexes are combined, and any further change of base can be made. This process is known as "splicing." [1]

[1] If the indexes overlap by two or more years, it might be advisable to total each index in the overlapping years (or in the most representative years), and to find the ratio of one sum to the other. This ratio could then be used as a common multiplier or divisor to equalize the two sums. There would probably remain some disagreement in the two indexes for the overlapping years, but this could be eliminated by averaging the two figures for each such year. For example, suppose that two abbreviated indexes in the same field, overlapping in the years 1928 and 1929, appear as follows

Year	Index A	Index B
1927	60	..
1928	61	83
1929	65	90
1930	..	81

In combining them, the ratio $\dfrac{83 + 90}{61 + 65} = 1.373$ indicates that index B, in these two years, is 1.373 times as great as index A. If they are combined upon the base of index A, index B must be divided by this factor. If the new base is to be that of index B, then index A must be multiplied by 1.373. An average of the conflicting items in each overlapping year is taken as the final index number, as follows:

Year	Index A × 1.373	Index B	"Spliced"
1927	82.4	..	82.4
1928	83.8	83	83.4
1929	89.2	90	89.6
1930	....	81	81.0

The same process of averaging is necessary if the combined index is calculated on the base of index A, in which event index B for each year is divided by 1.373. (This method sometimes becomes very complicated when several overlapping indexes are combined, but the procedure is as indicated, except that the geometric mean may be preferred.)

Deflating a value index.—If, in a given statistical field, an index of value (actual prices times actual quantities, aggregated and compared) and an index of prices are available, an index of quantity may obviously be obtained by the simple process of dividing the former index by the latter ($V \div P = Q$).

In practice, this principle, which is so clear in the abstract, becomes somewhat obscure because of the difficulty of obtaining indexes that are exactly comparable. However, when suitable indexes are obtainable, the process is comparatively simple. For example, an index of money wages may be divided by an index of the cost of living in order to measure changes in so-called real wages, that is, in the quantity of goods wages will buy. As an illustration the following figures have been selected from a certain industrial company:

Year	Wage per Week	Cost of Living
1926	$25.50	100.0
1927	26.20	99.1
1928	26.50	96.8
1929	26.90	96.7
1930	26.60	95.0

In order to reduce money wages to real wages it may be advisable first to reduce them to an index having the same base as the cost-of-living index. After this has been done, the wage index thus obtained may be divided by the cost-of-living index. The resulting quotients constitute an index of real wages, as follows:

Year	Indexes		
	Money Wages	*Cost of Living*	*Real Wages*
1926	100.0	100.0	100.0
1927	102.7	99.1	103.6
1928	103.9	96.8	107.3
1929	105.5	96.7	109.1
1930	104.3	95.0	109.8

These figures indicate that in this particular group, real wages—that is, the quantity of goods which money wages would buy—were rising during the period in question. Money wages, however, accounted for only a part of this rise, a large part of the change arising out of the decline in the cost of living (cf. Fig. 8·2, page 180).

It is not essential, of course, that wage figures be first

reduced to an index. If, for example, money wages had been divided by the cost-of-living index, weekly wages thus deflated would have expressed the change in real wages just as accurately. Thus expressed as deflated dollars, they could be designated as wages in terms of 1926 purchasing power and could have been reduced to an index if desired. Again, money wages are sometimes divided by wholesale prices to compare them with other manufacturing costs. Sometimes, also, values in special fields are divided by a general price index in order to remove the effect of the changing value of the dollar. The resulting index may then be described as a value index with variations in the purchasing power of the dollar in general markets removed. Broadly speaking, any aggregates or indexes, no matter what the base, if they express correctly the relative change in value exchanged and prices, may be thus divided ($V \div P$) to measure relative changes in quantities.[1]

The process of deflation thus described has many practical uses in connection with the measurement of business trends. For example, series of data such as bank debits, which represent the changing volume of check transactions, may be deflated in order to suggest the physical volume of business. Similarly, export figures, commonly quoted in dollars, may be deflated to express physical volume of exports, although if this is done it may be difficult to obtain a suitable price index. Many similar situations arise in the course of statistical analysis.[2]

Reference has been made in preceding paragraphs to the use of an "appropriate" index in deflating, and it may be worth while to express a word of caution in this connection.[3] Such an

---

[1] Use of the term "deflation" in this connection deserves brief explanation. The term came into use during the first World War when dollar wages and other values were rapidly rising. When these values were divided by an appropriate price index, the apparent rise was naturally reduced, hence they were said to be deflated. The term was then extended to cover other cases where a value index or an actual value is divided by a price index.

[2] A special case of deflation is $1 \div P$, which indicates changes in the purchasing power of the dollar, in the markets represented by P.

[3] Indexes of payrolls (see Fig. 8·4) are often used to suggest changes in business activity and are sometimes divided by an employment index. The quotient may then be deflated by a cost-of-living index. This practice is objectionable because in various stages of the business cycle the composition of the labor force with respect to higher- and lower-paid workers changes materially.
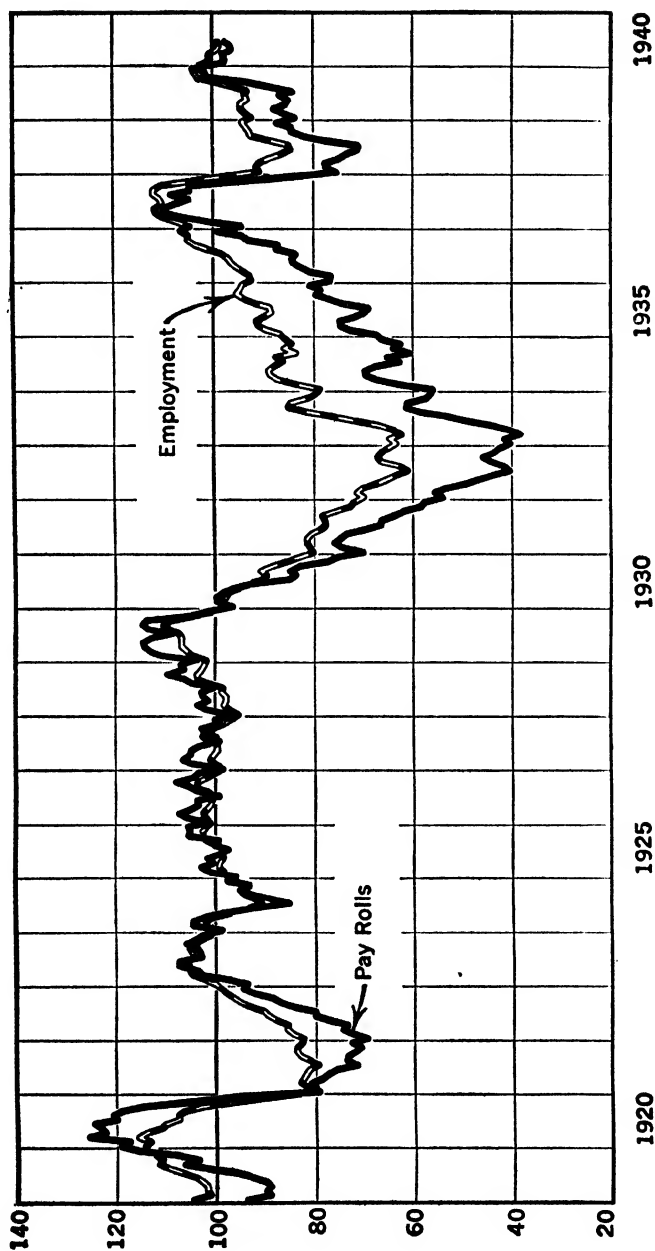
Fig. 8·4.—Indexes of Employment and Payrolls in Manufacturing Industries, 1919–1940. (1923–25 = 100.) Source: *Monthly Labor Review*, April, 1940, p. 1006.

index must be appropriate in that it is relevant to the data it is used to deflate, and it must be appropriate in that the adjustment it makes represents the sort of deflation that is assumed. For instance, in deflating money wages to obtain a measure of real wages, it is highly important that the deflating index represent living costs rather than wholesale prices, or farm prices, or the prices of foods, rents, or some other individual item or group of items within the general scope of living costs.   Moreover, it is important that the cost-of-living index represent such costs for the group whose wages are being deflated; that it is not limited to some part of them.   It is not uncommon, for instance, to find professional salaries deflated by the current indexes of living costs, most of which refer to laboring families.   Obviously, such practice is objectionable, since the budgets of the two groups are not closely comparable.   Again, such income may be deflated by an index that refers to but one locality and that not the one involved.   To be appropriate, an index must avoid all such deficiencies.[1]

For these and related reasons, cost-of-living indexes must be used with special caution.   They may readily become misleading as a result of shifts in consumption.   Until recently, the most widely used of these indexes were based primarily on budget studies made at the time of the first World War, although it is apparent that consumption habits in foods, clothing, amusements, transportation, house furnishings, and other major budget items have undergone a long series of significant changes in the past twenty years.   In recognition of this fact, some changes in methods of calculation of the Bureau of Labor Statistics index were introduced in 1935, and, in 1940, the Bureau announced a comprehensive new index based on average costs in the 1935–1939 period.[2]   However, no satisfactory

[1] There are two commonly used, nation-wide cost-of-living indexes for the United States, one of which is prepared by the National Industrial Conference Board (monthly) and the other of which is published by the Bureau of Labor Statistics. Both series have been based upon budget studies that are now antiquated.   (See Dale Yoder, *Labor Economics and Labor Problems*, New York, McGraw-Hill Book Co., 1939, pp. 250–254.)

[2] For an extended discussion of the method of calculating the new index, see "The Bureau of Labor Statistics' New Index of Cost of Living," *Monthly Labor Review*, 51 (2), August. 1940, pp. 367–404.

method of keeping such an index up to date has been devised.

Interpolating.—Many problems arise in the process of gathering, editing, and analyzing data for index numbers that are difficult to forecast and for which no exact rules of procedure can be given.   Sometimes, for example, it is necessary to interpolate a missing item.   To illustrate, it may be assumed that monthly. figures on bank debits in a certain area are being compiled from 1925 to date, but that for one city the data are lacking for August, 1935, and that such figures are not obtainable.   Perhaps it is noted that, in other years, the August figure is normally 5 per cent below the average of contiguous July and September figures.   If this rule seems to be constant, then the missing item might be supplied by taking 95 per cent of the average of July and September, 1935.   If, however, there appears to be no definite seasonal change the interpolation might be made on the basis of other comparable figures.   Perhaps bank clearings could be obtained for the given city.   Inspection would be likely to show that bank clearings, though a little smaller than bank debits, run a somewhat parallel course.   Suppose that bank clearings in August, 1935, were found to be 5 per cent below the average of July and September clearings for the same year.   It might then be fair to assume that the August debits should likewise be 5 per cent below their own July and September average, and an estimate could be prepared on this assumption.   In a similar manner local figures might be interpolated from state or national figures, assuming that they ordinarily run a parallel course.

If interpolation cannot be made by reference to other comparable data, it may sometimes be made on the assumption that the figures for a short period tend to follow a straight line or a simple curve such as a parabola, so that the interpolation may be made by reference to a calculated trend, as will be explained in the next chapter.

It should be obvious that no general rule can be made for such interpolations, nor is it possible to state to what extent they may be applied.   Indexes based upon these estimated data must necessarily lose in reliability.   Hence, in presenting

such indexes, the statistician will be careful to make clear the nature of the estimates. Good judgment and experience are essential in all such adjustments.

READINGS

(See also special and general references, pages 591 and 597.)

BOWLEY, A. L., "A New Index-Number of Wages," *Royal Economic Society*, Memorandum 12, January, 1929, pp. 1–7.

BURNS, ARTHUR F., "The Measurement of the Physical Volume of Production," *Quarterly Journal of Economics*, 44 (2), February, 1930, pp. 242–262.

COLE, ARTHUR H., and FRICKEY, EDWIN, "The Course of Stock Prices, 1825–1866," *Review of Economic Statistics*, 10 (3), August, 1928, pp. 117–139.

CUTTS, JESSE M., and DENNIS, SAMUEL J., "Revised Method of Calculation of the B.L.S. Wholesale Price Index," *Journal of the American Statistical Association*, 32 (200), December, 1937, pp. 663–674.

DAVENPORT, DONALD H., and SCOTT, FRANCES V., *An Index to Business Indices*, Chicago, Business Publications, Inc., 1937.

DENNIS, SAMUEL T., "The Sensitive Price Index," *Review of Economic Statistics*, 14 (1), February, 1932, pp. 42–44.

FISHER, IRVING, *The Making of Index Numbers*, Boston, Houghton Mifflin Co., 1927.

——"Weekly Wholesale Index," *Journal of the American Statistical Association*, 23 (164), December, 1928, pp. 429–434.

——"Wholesale Commodity Price Indexes," *Journal of the American Statistical Association*, 25 (171), September, 1930, pp. 308–315.

JEROME, HARRY, "Production," *American Journal of Sociology*, 34 (6), May, 1929, pp. 994–1002.

JOHNSON, NORRIS O., "New Indexes of Production and Trade," *Journal of the American Statistical Association*, 33, June, 1938, pp. 341–348.

LEONG, Y. S., "Indexes of the Physical Volume Production of Producers' and Consumers' Goods," *Journal of the American Statistical Association*, 27 (177) March, 1932, pp. 21–36.

——"Indexes of the Physical Volume Production of Producers' Goods, Consumers' Goods, Durable Goods, and Transient Goods," *Journal of the American Statistical Association*, 30 (190) June, 1935, pp. 361–377.

MAXWELL, FLOYD W., "The Revised Index of the Volume of Manufacture," *Review of Economic Statistics*, 11 (2) May, 1929; pp. 68–109.

MITCHELL, W. C., "The Making and Using of Index Numbers," United States Bureau of Labor Statistics, *Bulletin 284*, 1921, or *Bulletin 656*.

"New Federal Reserve Index of Industrial Production," *Federal Reserve Bulletin*, 26 (8) August, 1940, pp. 753–771.

PERSONS, W. M., *The Construction of Index Numbers*, Boston, Houghton Mifflin Co., 1928.

"Revision of Federal Reserve Board's Index of Industrial Production," *Survey of Current Business*, 20 (8), August, 1940, pp. 11–18.

WARREN, GEORGE F., and PEARSON, FRANK A., "A Monthly Index Number of Wholesale Prices in the United States for 135 Years," *Proceedings of the American Statistical Association*, 26 (173A), March, 1931, pp. 244–249.

"Wholesale Prices," *Bulletins* of the United States Bureau of Labor Statistics, Government Printing Office, Washington.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Use the following data to compute index numbers of value, quantity, and price by the common aggregative method, base-weighted (1933 = 100).

	1933			1934			1935			1936			1937		
	q	p	v	q	p	v	q	p	v	q	p	v	q	p	v
Commodity A	5 @ 4	=	20	6	3		4	5		6	6		5	5	
Commodity B	12 @ 2	=	24	15	3		15	6		18	7		12	6	
Commodity C	3 @ 2	=	6	6	2		6	4		6	5		3	1	

Σv 50
Index   100

2. From the following four series of data representing quantities and prices, compute index numbers of value, quantity, and price, selecting weights from the first year, taken as the base.

(a)

	1926		1927		1928		1929	
	q	p	q	p	q	p	q	p
Commodity A	2	3	4	3	3	4	2	6
Commodity B	3	5	6	3	6	4	5	6
Commodity C	2	2	3	3	5	1	4	2

(b)

	1926		1927		1928		1929	
	q	p	q	p	q	p	q	p
Commodity A	4	1	6	3	8	5	2	4
Commodity B	5	2	6	3	7	2	5	6
Commodity C	3	2	6	1	6	3	3	4

(c)

	1926		1927		1928		1929	
	q	p	q	p	q	p	q	p
Commodity A	1	2	2	2.5	1.5	4	2	2
Commodity B	2	2	3	2	2	3	4	2
Commodity C	1	4	1.5	4	2	3	2	4

(d)

	1935		1936		1937		1938	
	q	p	q	p	q	p	q	p
Commodity A	2	2	3	3	2	1.5	4	1
Commodity B	4	2	6	2	6	3	8	1
Commodity C	8	1	6	1.5	4	2	4	2

3. The data summarized below roughly approximate annual production and prices of three important commodities in the United States for the years 1890–1914. (Quantities in millions, prices in dollars.)

(a) Compute base-weighted aggregative indexes of value, quantity, and price for the combined series.

(b) Recompute the quantity index using median prices in 1910–1914 as weights, namely 3, 2, and 2, for A, B, and C, respectively, and making the average of 1910–1914 the base.

(c) Change the base of the price index obtained in (a) so that the average for 1910–1914 is 100; that is, divide P as obtained in (a) by $(122 + 112 + 132 + 132 + 90) \div 5 = 117.6$. Note that the results will be the same as if computed as in (b), since quantities in 1890 (5; 10; 10) are in the same ratios as the medians of 1910–1914 (10; 20; 20).

Year	Commodities					
	A		B		C	
	q	p	q	p	q	p
(Base) 1890	5	4	10	2	10	1
1891	6	4	11	1	11	2
1892	5	2.8	10	1.6	10	1.5
1893	4	2	7	2	9	1
1894	4	3	8	1.5	10	0.6
1895	5	2.6	11	2	10	0.4
1896	6	3	12	1.5	10	0.2
1897	6	2	13	1	12	1.5
1898	7	3	14	1	10	1.3
1899	6	3	11	2	15	1
1900	7	4	14	1.5	12	1.5
1901	6	5	12	1	13	1
1902	8	5	17	2	15	1
1903	7	4	15	2	16	1.5
1904	8	5	16	1.5	15	2
1905	9	4	17	2	18	1.5
1906	10	3.4	18	2	21	2
1907	8	5	20	1.4	19	2
1908	9	3	16	3	16	1
1909	9	4	20	1.9	17	2
1910	10	3	20	2.1	18	2.5
1911	9	2	19	2	20	2.6
1912	10	5.2	21	2	23	2
1913	10	5.2	21	2	22	2
1914	9	2	18	1.5	20	2

1.

YEAR	V	Q	P
1933	100	100	100
1934	150	132	114
1935	268	116	218
1936	384	144	258
1937	200	100	200

2. (a)

YEAR	V	Q	P
1926	100	100	100
1927	156	192	84
1928	164	196	88
1929	200	156	136

(b)

YEAR	V	Q	P
1926	100	100	100
1927	210	150	150
1928	360	170	195
1929	250	90	290

(c)

YEAR	V	Q	P
1926	100	100	100
1927	170	160	105
1928	180	150	130
1929	200	200	100

(d)

YEAR	V	Q	P
1935	100	100	100
1936	150	120	130
1937	145	100	155
1938	100	140	110

3. (a) Indexes for successive years, 1890–1914.

V $\dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$	Base-weighted		V $\dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$	Base-weighted		V $\dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$	Base-weighted	
	Q $\dfrac{\Sigma p_0 q_1}{\Sigma p_0 q_0}$	P $\dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$		Q $\dfrac{\Sigma p_0 q_1}{\Sigma p_0 q_0}$	P $\dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$		Q $\dfrac{\Sigma p_0 q_1}{\Sigma p_0 q_0}$	P $\dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$
100	100	100	110	122	90	182	168	110
114	114	100	134	136	100	216	186	118
90	100	90	110	122	90	234	196	122
62	78	80	178	162	110	216	188	112
60	84	72	164	148	110	280	210	132
78	104	74	188	158	120	276	208	132
76	116	64	194	176	110	170	184	90
86	124	70	224	194	114			
96	132	76	212	182	118			

(b)

YEAR	Q	YEAR	Q	YEAR	Q
1890	50.2	1899	63.9	1907	93.1
1891	56.6	1900	66.6	1908	83.0
1892	50.2	1901	62.0	1909	92.2
1893	40.1	1902	80.3	1910	96.7
1894	43.8	1903	75.7	1911	95.8
1895	52.0	1904	78.5	1912	107.7
1896	56.6	1905	88.5	1913	105.8
1897	62.0	1906	98.5	1914	94.0
1898	63.0				

(c)

YEAR	P	YEAR	P	YEAR	P
1890	85	1899	77	1907	100
1891	85	1900	85	1908	94
1892	77	1901	77	1909	100
1893	68	1902	94	1910	104
1894	61	1903	94	1911	95
1895	63	1904	102	1912	112
1896	54	1905	94	1913	112
1897	60	1906	97	1914	77
1898	65				

B. PROBLEMS

4. Prepare a series of index numbers representing the annual volume of passenger car sales by General Motors from 1929 to 1939 inclusive, using the data of Problem 5, Chapter IV, page 82. Use the year 1929 as a base.

5. Use the following table of annual prices of commodities as marketed by Iowa farmers for the years indicated, together with quantity weights based on the years 1925–1929, to calculate price indexes for these years, by the common aggregative method.

Years	Hogs (cwt.)	Cattle (cwt.)	Sheep (cwt.)	Corn (bu.)	Oats (bu.)	Wheat (bu.)	Hay (ton)	Butter (lb.)	Eggs (doz.)	Poultry (lb.)
1910-14	7.304	6.388	4.510	0.528	0.345	0.850	9.822	0.254	0.169	0.098
1915-19	12.562	9.668	8.022	1.074	0.540	1.668	13.648	0.376	0.278	0.158
1920-24	8.680	7.650	6.052	0.718	0.402	1.244	12.658	0.410	0.262	0.176
1925	11.00	8.49	7.46	0.90	0.38	1.41	11.23	0.41	0.260	0.180
1926	11.60	7.99	6.89	0.61	0.34	1.28	13.97	0.42	0.260	0.197
1927	9.54	8.97	6.56	0.74	0.41	1.22	13.38	0.44	0.210	0.180
1928	8.55	10.90	6.96	0.81	0.42	1.08	12.04	0.46	0.250	0.200
1929	9.41	10.80	6.52	0.78	0.39	1.07	11.44	0.46	0.260	0.194
1930	8.80	9.17	4.67	0.70	0.33	0.77	9.31	0.36	0.190	0.154
1931	5.64	6.50	2.88	0.43	0.21	0.44	8.29	0.27	0.140	0.137
1932	3.20	4.95	1.98	0.23	0.15	0.37	7.56	0.20	0.107	0.095
1933	3.33	4.34	2.25	0.27	0.22	0.69	5.36	0.21	0.106	0.075
1934	4.15	4.96	2.78	0.57	0.39	0.88	11.16	0.24	0.130	0.103
1935	8.66	8.16	3.81	0.72	0.32	0.88	10.95	0.29	0.207	0.143
1936	9.30	7.34	3.70	0.76	0.34	1.03	8.80	0.33	0.181	0.137
1937	9.59	9.05	4.06	0.90	0.34	1.03	10.04	0.34	0.180	0.157
1938	7.68	7.79	3.26	0.41	0.20	0.61	6.57	0.28	0.153	0.131
Quantity weights:	6.388	2.766	0.157	19.410	12.877	1.453	0.078	39.778	30.956	29.828

6. The data summarized herewith represent index numbers of money wages and living costs (for manufacturing workers) for the years from 1913 to 1934 inclusive. Deflate the money wage index to secure a series of indexes of real wages for these years.

INDEX NUMBERS OF WAGES PER HOUR AND COST OF LIVING

Year	Index numbers of	
	Wages per hour (money wages)	Cost of living
1913	100	100.0
1914	102	103.0
1915	103	105.1
1916	111	118.3
1917	128	142.4
1918	162	174.4
1919	184	188.3
1920	234	208.5
1921	218	177.3
1922	208	167.3
1923	217	171.0
1924	223	170.7
1925	226	175.7
1926	229	175.2
1927	231	172.7
1928	232	170.7
1929	233	170.8
1930	229	163.7
1931	217	148.1
1932	186	133.9
1933	178	131.7
1934	200	137.7

7. On the basis of the following wage (factory average weekly earnings, United States) and cost-of-living data, construct an index of real wages, taking 1932 as the base year.

Year	Wages	Cost of living
1932	$18.12	77.9
1933	17.57	74.9
1934	19.14	79.4
1935	21.06	82.6
1936	22.82	84.8
1937	25.14	88.5
1938	22.83	86.4

8. Combine the two wholesale price indexes summarized below into a single series having 1890 as the base year:

	Year		Index *A*	Index *B*
(Base *A*)	1890	.	100	
	. . . .		. . .	
	1911		150	
	1912		147	98
Base (*B*)	1913			100
	1914			110
	1915			108

9. The following two indexes of real wages for the same types of workers living in the same industrial area are available to the management of a concern:

Year	Index *A* 1914 = 100	Index *B* 1913 = 100
1924	135	
1925	143	
1926	144	130.7
1927	140	133.8
1928		135.9
1929		136.4
1930		139.9
1931		146.5
1932		138.9

(*a*) Link the two indexes to provide one continuous index with 1914 as a base.

(*b*) Link the two indexes to provide one continuous index with 1913 as a base.

(*c*) Shift the base of the continuous index to 1929.

10. The Bureau of Labor Statistics is now presenting intercity comparisons of living costs from which the data below were selected. The data are for 1935 and are expressed as percentages of costs in Washington, D. C.

Recalculate the relatives for each series, using the city most representative of your vicinity as a base.

City	Total	Food	Clothing	Housing	Fuel, etc.	Misc.
Washington	100	100	100	100	100	100
Minneapolis	98	92	111	77	134	106
Los Angeles	92	93	115	58	104	115
Butte	90	94	120	61	122	84
Buffalo	89	93	103	61	100	101
Denver	88	91	102	60	94	105
Seattle	87	93	108	49	109	98
Memphis	86	91	97	65	87	96
Dallas	84	95	90	63	84	86
Mobile	79	91	92	48	94	84

11. Given the following data, representing food prices in a middle-western industrial area, calculate indexes of retail food prices for May, 1934, and February, 1939 (1926 = 100), using the common aggregative method.

Item	Unit	Annual family con-sumption	Prices (in cents)		
			1926	May, 1934	Feb., 1939
Sirloin	Pound	34	32.0	27.1	39.0
Pork chops	Pound	45	35.5	24.1	29.4
Leg of lamb	Pound	2	35.5	26.5	27.9
Milk, fresh	Quart	364	11.0	9.0	12.2
Butter	Pound	53	49.8	28.5	33.0
Eggs	Dozen	53	41.0	21.4	29.9
Bread	Pound	521	9.5	8.1	8.0
Corn	No. 2 can	13	15.2	9.6	10.9
Sugar	Pound	154	7.1	5.3	5.1
Tea	Pound	5	61.3	59.3	70.4
Raisins	Pound	11	15.1	10.0	10.0
Potatoes	Pound	810	4.1	2.5	2.4
Coffee	Pound	45	54.0	29.2	22.8

12. The following indexes of value, quantity, and production in the United States were compiled by the Bureau of Foreign and Domestic Commerce.

Plot the three primary indexes (columns 5, 6, and 7) on semi-logarithmic paper.

(For a discussion of the significance of these figures and the method of compilation, see the *Survey of Current Business,* May, 1936, and September, 1939.)

ESTIMATED AGGREGATE VALUE AND PHYSICAL VOLUME OF GOODS MARKETED
AT WHOLESALE IN THE UNITED STATES, 1899–1938

	1	2	3	4	5	6	7
Year	Aggregate value index (1929 = 100)	Aggregate value of domestic production (millions of dollars)	Imports for consumption including duties paid (millions of dollars)	Total value of goods marketed at wholesale	Index of value of goods marketed at wholesale (1929 = 100)	Index of wholesale prices (1929 = 100)	Index of physical volume of goods marketed at wholesale (1929 = 100)
				(2+3)			(5÷6)
1899	17.9	14,137	888	15,025	17.9	54.8	32.7
1900	19.2	15,163	1,060	16,223	19.3	58.9	32.8
1901	19.1	15,084	1,042	16,126	19.2	58.0	33.1
1902	23.3	18,401	1,151	19,552	23.3	61.8	37.7
1903	22.9	18,086	1,289	19,375	23.1	62.5	37.0
1904	23.1	18,243	1,240	19,483	23.2	62.6	37.1
1905	26.0	20,534	1,345	21,879	26.1	63.1	41.4
1906	28.7	22,666	1,507	24,173	28.8	64.8	44.4
1907	30.1	23,772	1,744	25,516	30.4	68.4	44.4
1908	27.8	21,955	1,466	23,421	27.9	66.0	42.3
1909	32.6	25,746	1,577	27,323	32.6	70.9	46.0
1910	35.1	27,721	1,874	29,595	35.3	73.9	47.8
1911	31.6	34,956	1,838	26,794	31.9	68.1	46.8
1912	38.2	30,169	1,946	32,115	38.3	72.5	52.8
1913	37.7	29,774	2,080	31,854	38.0	73.2	51.9
1914	37.5	29,616	2,190	31,806	37.9	71.5	53.0
1915	44.1	34,828	1,975	36,803	43.9	72.9	60.2
1916	57.8	45,648	2,573	48,221	57.5	89.7	64.1
1917	87.5	69,104	3,124	72,228	86.1	123.3	69.8
1918	94.3	74,474	3,123	77,597	92.5	137.8	67.1
1919	94.7	74,790	4,065	78,855	94.0	145.4	64.6
1920	117.1	92,480	5,428	97,908	116.7	162.0	68.9
1921	64.3	50,782	2,849	53,631	63.9	102.4	62.4
1922	75.0	59,232	3,525	62,757	74.8	101.5	73.7
1923	87.9	69,420	4,299	73,719	87.9	105.6	83.2
1924	82.5	65,155	4,107	69,262	82.6	102.9	80.3
1925	91.0	71,868	4,728	76,596	91.3	108.6	84.1
1926	94.5	74,632	4,998	79,630	94.9	104.9	90.5
1927	90.5	71,473	4,738	76,211	90.8	100.1	90.7
1928	97.1	76,686	4,620	81,306	96.9	101.5	95.5
1929	100.0	78,976	4,924	83,900	100.0	100.0	100.0
1930	78.5	61,996	3,576	65,572	78.2	90.7	86.2
1931	57.8	45,625	2,459	48,084	57.3	76.6	74.8
1932	42.7	33,723	1,584	35,307	42.1	68.0	61.9
1933	45.0	35,576	1,717	37,293	44.4	69.2	64.2
1934	54.3	42,884	1,937	44,821	53.4	78.6	67.9
1935	65.1	51,424	2,396	53,820	64.1	83.9	76.4
1936	77.0	60,812	2,832	63,644	75.9	84.8	89.5
1937	87.4	69,073	3,480	72,553	86.5	90.6	95.5
1938	73.2	57,810	2,251	60,061	71.6	82.5	86.8

CHAPTER IX

INDEX NUMBERS (*Continued*)

In the preceding chapter, attention was directed to the most commonly used methods of constructing index numbers. Little consideration was there given to the mathematical validity of such relatives. As a matter of fact, however, it should be noted that, when these methods and their results are critically appraised, they prove to be useful approximations rather than exact measures. Indeed, it can be demonstrated that there are certain innate inconsistencies in most of the commonly used indexes, and there are also methods by means of which these shortcomings may be removed or reduced. In this chapter, therefore, attention is turned to the most important of these inconsistencies and to the adjustments or refinements they suggest as desirable.

Fisher's " ideal " index formula.—Professor Irving Fisher has criticized the more common methods of constructing index numbers on the ground that the use of a single weight may occasion an inconsistency or discrepancy in a series of index numbers. He has suggested a way out of this difficulty when adequate data are available, and attention may well be given to his procedure.

In order to make use of Fisher's method, it is necessary to have both quantity and price data for the period under consideration. The method then proceeds to the calculation of index numbers of quantity, of price, and of the product of quantity and price, that is, value. The value index is made up in a manner similar to that used in the common aggregative method, except that the aggregates are the sum of actual prices times actual quantities, *no weights being required*. The resulting aggregates, therefore, reflect a combination of price and quan-

tity changes; that is, the income derived from the recorded sales. In constructing the price and quantity indexes, the weights first employed are selected from the period chosen as a base. Hence the resulting indexes are spoken of as *base-weighted* (P_b and Q_b). However, weights selected from each given year are indirectly employed as is explained later. The entire computation of the value index and of aggregative base-weighted price and quantity indexes is illustrated with simple data in Example 9·1, part I.

The essential feature of Fisher's method is a correction that is applied to these base-weighted indexes. That some correction is desirable is suggested by the fact that, for any given year, the product of that year's price index and the same year's quantity index is usually different from the value index for that year. Theoretically, the product should equal the value index, just as any price multiplied by the quantity sold should represent the value. This comparison is known as the *factor's test*.

In the Fisher procedure (see part II of Example 9·1), this test is applied by dividing each year's value index by that year's price index ($V \div P_b = Q_r$) to see whether it checks with the given quantity index. If it does, the base-weighted index is taken as final, but if it does not, the result of the division is taken as a second estimate of the quantity index and is set down in a separate column (Q_r). A second estimate of the price index is similarly found by dividing the year's value index by its quantity index ($V \div Q_b = P_r$). The revised indexes thus obtained are denoted as *reverse-weighted* indexes, because they are identical with the results that would be obtained if the indexes were recalculated with weights chosen from the given year instead of the base year.

After the second estimates of P and Q have been obtained for each year in the series, the two estimates are averaged to obtain the final result. That is, according to Fisher's method,

$$P = \sqrt{P_b \times P_r}, \quad \text{or, approximately,} \quad (P_b + P_r) \div 2$$

$$Q = \sqrt{Q_b \times Q_r}, \quad \text{or, approximately,} \quad (Q_b + Q_r) \div 2$$

EXAMPLE 9·1

FISHER'S IDEAL METHOD

Data: Assumed price and quantity data, 1926–1929, in dollars per unit

I. Value index, with price and quantity base-weighted indexes

Commodities	Base = 1926 p	q	v	1927 p	q	v	1928 p	q	v	1929 p	q	v
A (bushels)	5	40	200	4	35	140	4	45	180	5	50	250
B (pounds)	3	20	60	5	25	125	4	30	120	2	40	80
C (feet)	8	30	240	7	40	280	8	30	240	6	45	270
Σpq			500			545			540			600
Value index			100			109			108			120

	p	q_w	v	p	q_w	v	p	q_w	v	p	q_w	v
A (bushels)	5	40	200	4	40	160	4	40	160	5	40	200
B (pounds)	3	20	60	5	20	100	4	20	80	2	20	40
C (feet)	8	30	240	7	30	210	8	30	240	6	30	180
Σpq_w			500			470			480			420
Price Index			100			94			96			84

	q	p_w	v	q	p_w	v	q	p_w	v	q	p_w	v
A (bushels)	40	5	200	35	5	175	45	5	225	50	5	250
B (pounds)	20	3	60	25	3	75	30	3	90	40	3	120
C (feet)	30	8	240	40	8	320	30	8	240	45	8	360
Σqp_w			500			570			555			730
Quantity index			100			114			111			146

II. Recalculation

Year	Indexes: see Part I V	P_b	Q_b	$V \div Q_b$ P_r	$V \div P_b$ Q_r	$(P_b + P_r) \div 2$ P	$(Q_b + Q_r) \div 2$ Q
1926	100	100	100	100.0	100.0	100.0	100.0
1927	109	94	114	95.6	116.0	94.8	115.0
1928	108	96	111	97.3	112.5	96.6	111.8
1929	120	84	146	82.2	142.9	83.1	144.4

Strictly speaking, the geometric mean should be employed instead of the arithmetic mean in this final averaging, but in practice the latter is usually taken as a reasonably close approximation. When the geometric mean is used, P times Q will exactly equal V and the factor's test is satisfied.[1]

Fisher's method of computing index numbers is not in general use, principally because of the cost involved in collecting adequate data. For example, if an index of wholesale prices in the United States were computed by this method it would be necessary to gather, each week, not only the prices of the 813 commodities involved, but also the quantities marketed as well. This would presumably improve in some small degree the accuracy of the index, but this advantage would probably not warrant the additional labor and expense. Besides, although the method removes an inconsistency, it is by no means above criticism from a theoretical point of view.

Weighted-relatives method.—Since the earliest development of index numbers, question has frequently been raised as to the aggregative process. As an alternative, another method utilizing a somewhat different approach is frequently preferred. As a first step, it calculates abstract "relatives" or simple index numbers. For example, in an index of wholesale prices, if a commodity in the base year, 1926, cost $0.50 a bushel and in 1939 cost $0.40 a bushel, the latter price would be indicated by an index number, or relative, of 80 (per cent). In the same way, each commodity would be expressed as a percentage relative to the corresponding price in the base year. One thing that may be said in favor of the method is that it affords a convenient means of comparing price changes among different commodities.

The problem next presented in calculation of a composite price index of this type is the averaging of the relatives for any given period. Obviously weights must be used, and at first thought it might be assumed that typical quantities should be taken as the weights. But such weights would be unsatisfactory

[1] The two aggregative price formulas combined in Fisher's index, namely, the base-weighted and the given-weighted, are often referred to in the literature of index numbers as the beta and gamma formulas, respectively. They are also designated as Laspeyre's and Paasche's indexes, respectively, after the scientists who first popularized them.

because the resulting average would be greatly affected by the kind of unit employed.[1]   For example, if one of the commodities is iron, the quantity weight will be increased 2,000 times if the

<div align="center">EXAMPLE 9·2</div>

<div align="center">THE RELATIVE METHOD</div>

Data: See Example 9·1.

I. Using arithmetic mean

(1) Base prices, 1926, p_0	(2) Given prices, 1927, p_1	(3) Price relatives, p_1/p_0	(4) Base weights, v_0 (value in base year)	(5) Product (3) × (4) $v_0(p_1/p_0)$
A (bushels) 5	4	80	200	16,000
B (pounds) 3	5	$166\frac{2}{3}$	60	10,000
C (feet) 8	7	$87\frac{1}{2}$	240	21,000

$$\Sigma = 500 \qquad 500\overline{)47{,}000}$$

Index, 1926 = 100   Index, 1927 = 94 (%)

II. Using geometric mean

(1) Base prices, p_0	(2) Given prices, p_1	(3) Price relatives, p_1/p_0	(4) Log of relatives, $\log (p_1/p_0)$	(5) Base weights, v_0	(6) Product (4) × (5) $v_0 \times \log$
A (bushels) 5	4	80	1.90309	200	380.61800
B (pounds) 3	5	$166\frac{2}{3}$	2.22185	60	133.31100
C (feet) 8	7	$87\frac{1}{2}$	1.94201	240	466.08240

$$500\overline{)980.01140}$$

Logarithm of the index   1.96002

Index, 1926 = 100   Index, 1927 = 91.21 (%)

---

[1] Such weights violate the so-called *units test,* mentioned later, which requires that an index-number formula should result in an index which does not vary if the formula is again applied to the same data stated in different physical units.

price is quoted per pound over what it would be if it were quoted per short ton, and consequently the final average would be subject to change.   In the aggregative method, on the contrary, a change in the unit has no final effect on the index.   Hence, *values* are taken as weights, since they not only avoid this difficulty, but at the same time reflect the relative importance of given commodities in the market.   They therefore meet the test of validity known as the *units test*.   The process of obtaining such a weighted average is illustrated for simple assumed data in Example 9·2, and for data of Iowa prices of farm products in Example 9·3.   In the former case selected weights are the values (price times quantity) in the base year, but any values regarded as typical might be used.   The use of the geometric mean in the second case illustrated in each example is discussed later in this chapter.

It will be seen that the use of the arithmetic mean in Example 9·2 results in exactly the same price index for 1927 as was previously obtained by the base-weighted aggregative method.   This will always be true when the value weights ($q \times p$) are chosen from the base year.[1]   Hence, when such weights are used there is little to be said in its favor, and it has few advantages.   This type of index, however, has been used in measuring changes in the cost of living, where separate indexes for food, clothing, rent, etc., are combined into a single composite index.   In such cases, the separate indexes are virtually relatives, and are combined by the use of weights representative of the values concerned.   By this method, one of the most widely used cost-of-living indexes (National Industrial Conference Board) is computed by combination of separate indexes for five classes of items, each of which is weighted according to its proportion in a

---

[1] The relative method applied to prices, base-weighted, is expressed by the formula,

$$P_1 = \sum\left(\frac{p_1}{p_0} \times p_0 q_0\right) \div \sum p_0 q_0$$

But p_0 may be canceled within the parenthesis following the first summation sign, giving

$$\Sigma p_1 q_0 \div \Sigma p_0 q_0$$

which is the formula for the base-weighted method.

EXAMPLE 9·3

PRICE INDEX, METHOD OF WEIGHTED RELATIVES, ARITHMETIC AND GEOMETRIC MEANS

Data: Prices received by Iowa farmers in base period, 1910–1914, and in February, 1940. Weights are percentages of value marketed in February, 1940. (Source: Iowa State College Bulletins, Ames, Iowa.)

Commodities	Arithmetic mean of relatives					Geometric mean of relatives		
	Prices ($)		(3)	(4)	(5)	(6)	(7)	(8)
	(1)	(2)	p_1/p_0 Relative	Weights, per cent value	Product $(p_1/p_0)w$	$\log(p_1/p_0)$	Weights, see col. (4)	Product
	p_0 1910–14	p_1 Feb. 1940						
Hogs (cwt.)	7.304	4.90	67.1	52.7	3,536.17	1.82672	52.7	96.268144
Cattle (cwt.)	6.388	·8.10	126.8	16.3	2,066.84	2.10312	16.3	34.280856
Sheep (cwt.)	4.510	4.85	107.5	0.7	75.25	2.03141	0.7	1.421987
Corn (bu.)	0.528	0.47	89.0	11.9	1,059.10	1.94939	11.9	23.19741
Oats (bu.)	0.346	0.36	104.0	2.3	239.20	2.01703	2.3	4.639169
Wheat (bu.)	0.852	0.87	102.1	0.6	61.26	2.00903	0.6	1.205418
Hay (ton)	9.822	6.80	69.2	0.6	41.52	1.84011	0.6	1.104066
Butter (lb.)	0.254	0.30	118.1	10.1	1,192.81	2.07225	10.1	20.929725
Eggs (doz.)	0.168	0.160	95.2	2.6	247.52	1.97864	2.6	5.144464
Poultry (lb.)	0.098	0.105	107.1	2.2	235.62	2.02979	2.2	4.465538
				100.0	8,755.29		100.0	192.657108
				Index:	87.6		Log index:	1.926571
							Index:	84.4

typical working-class budget as indicated by a study of such budgets.

These proportions are as follows:

Food	33 per cent
Housing	20
Clothing	12
Fuel and light	5
Sundries	30
Total	100 per cent

Correcting for bias.—The relative method utilizing the arithmetic mean is sometimes said to have an "upward bias," a defect that may be offset by using the geometric mean instead of the arithmetic. That an upward bias is present may be proved by reversing the base and recalculating the index number with the same weights. To illustrate, prices may be assumed, together with appropriate value weights, as follows:

COMMODITY	UNIT	PRICES		WEIGHTS (v)
		1926	1927	
A	pound	$0.10	$0.20	1
B	bushel	0.80	1.00	2

If the year 1926 is taken as a base, the price relatives are 200 and 125, respectively, which are averaged as follows:

	FORWARD INDEX (p_2/p_1)	WEIGHT (v)	PRODUCT
A	200	1	200
B	125	2	250
			3)450

Index, 1926 = 100   Index, 1927 = 150

If, however, 1927 is regarded as the base, the 1926 relatives become 50 and 80, respectively, and are averaged as follows:

	BACKWARD INDEX (p_1/p_2)	WEIGHT (v)	PRODUCT
A	50	1	50
B	80	2	160
			3)210

Index, 1927 = 100   Index, 1926 = 70

It will be observed that for each commodity, A and B, the relative "forward" multiplied by the relative "backward" necessarily equals unity. That is, 200 per cent × 50 per cent = 1, and 125 per cent × 80 per cent = 1. If the method is correct, the same should axiomatically be true of the two composite indexes just obtained, that is, the product 150 per cent × 70 per cent should equal 1. As a matter of fact, however, it equals 105 per cent. This product is characteristic; indeed, it will always be found, except when the price changes are uniformly alike, that the product of the two indexes thus obtained is too great. This is evidence of an "upward bias," and the test by which it is revealed is called the *base-reversal test*.

If, however, the geometric mean is employed, the upward bias disappears, as is illustrated by the accompanying computations:

	FORWARD INDEX (p_2/p_1)	LOG (p_2/p_1)	WEIGHT, (v)	PRODUCT
A	200	2.30103	1	2.30103
B	125	2.09691	2	4.19382
				3)6.49485
				2.16495

Index, 1926 = 100. Index, 1927 = 146.20.

	BACKWARD INDEX (p_1/p_2)	LOG (p_1/p_2)	WEIGHT (v)	PRODUCT
A	50	1.69897	1	1.69897
B	80	1.90309	2	3.80618
				3)5.50515
				1.83505

Index, 1927 = 100. Index, 1926 = 68.39.

It will be seen that, when the geometric mean is applied to the relatives, the "forward" index, 146.2 per cent, multiplied by the "backward" index, 68.4 per cent, equals unity. The relative method may, therefore, be corrected for its upward bias by substituting the geometric mean for the arithmetic mean. If the relative method is to be used, therefore, it is desirable that calculations be based on the geometric mean. The labor thus involved argues heavily in favor of the aggregative method,

which, with suitable weights, does not necessarily have an upward bias.

Tests.—The three methods of computing complex index numbers that have now been discussed, namely, the common aggregative, the relative and Fisher's "ideal" are all very useful to the statistician. The first two, particularly the aggregative, are most commonly used, while the last is valuable as a theoretical check.

Unweighted averages are sometimes employed, but there is little to be said in their favor even as crude approximations, except perhaps in certain cases where the selection of data is so arranged as practically to have the effect of weights. Of the three methods, Fisher's is, of course, assumed to give the most accurate representation of the data. It is, however, subject to the criticism that it averages prices, which are ratios, by the same method employed for quantities, which constitute one factor in determining price.[1]

The three criteria to which index numbers should approximately conform may be summarized as follows:

(1) Units test. A change in the physical unit (as from tons to pounds or yards to inches) should make no difference to the index as computed.

(2) Factors test. When value, quantity, and price indexes are computed from the same set of data, and having the same base, in any given time period the price index times the quantity index should equal the value index.

(3) The base-reversal test. On the basis of the same data and method, a "forward" index (as from 1926 as base to 1927, or from city A as base to city B) should be the reciprocal of the "backward" index (as from 1927 as base to 1926, or from city B as base to city A).

---

[1] Probably the best theoretical basis for index numbers of a market is one which follows the usages of Keynes and others. Quantities in a market can be combined only by reference to their valuation, hence the logical physical unit is a "dollar's worth" at a price taken as prevailing or typical. If quantities and prices are expressed in terms of these units, the theoretical problem discussed by Fisher disappears. Q, the quantity index, becomes a common aggregative, and P is price weighted by the quantities of a given period. Or, more conveniently, $P = V \div Q$. (See R. Frisch, "The Problem of Index Numbers," *Econometrica*, January, 1936, pp. 1–38.)

Other tests have been suggested, and it was formerly customary to apply the factors test in the form of a so-called "factors-reversal" test in which an interchange of the p's and q's were assumed to interchange P and Q. This test, however, has been found to be misleading, since prices are ratios, and quantities represent one of the elements of which the ratios are composed, and therefore are not necessarily subject to the same method of averaging.

As has been seen, it is not essential that methods in common use conform exactly to these three tests. They are, however, important as setting up theoretical standards for these approximations, and for the choice of methods when accurate and adequate data are at hand.

Present status of index numbers theory.—Statisticians have been so interested in the practical applications of index-number theory that they have given little attention to the fundamental principles involved. As has been seen, there is at present no perfectly valid method of computing index numbers. There are merely methods of approximation which come more or less close to the theoretical norm, but to which, in greater or less degree, theoretical objections can still be raised.

As the theory stands at present, it may be said that a price may be regarded as a physical unit designed to make commensurable, in market terms, quantities expressed in such units as pounds, yards, etc. To say that commodity A is worth x cents per pound is equivalent to saying that a market unit of a dollar's worth is the quantity $1/x$ pounds. The number of physical units thus determined may then be equated to similar measures in other commodities. Hence on the assumption of certain standard prices, quantities may be aggregated. Theoretically, a price index is the value of certain commodities at a stated time expressed as a ratio to the physical volume in market units. Consequently both a quantity and a price index assume certain standard prices, and vary according to the assumption made. Thus the problem reduces to one which fundamentally is as complex as the measurement of mass and motion in Einstein's physics. Fortunately, however, the practical phases of the subject can be pursued with useful results even though the

fundamental philosophy of the subject awaits a clearer understanding of the nature of the units dealt with in economic analysis.

READINGS

(See also special and general references, pages 591 and 597.)

DAVIES, GEORGE R., "Index Numbers in Mathematical Economics," *Journal of American Statistical Association*, Supplement 27 (177A), March, 1932, pp. 58–64.

FRICKEY, EDWIN, "Revision of the Index of General Business Conditions," *Review of Economic Statistics*, 14 (2), May 15, 1932, pp. 80–87.

FRISCH, RAGNAR, "Annual Survey of General Economic Theory: The Problem of Index Numbers," *Econometrica*, 4 (1), January, 1936.

HUHN, R. VON, "Relation between the Arithmetic Average and Geometric Average of Two Index Numbers," *Journal of the American Statistical Association*, 25 (169), March, 1930, pp. 76–79.

KONUS, A. A., "The Problem of a True Index of the Cost of Living," *Econometrica*, 7 (1), January, 1939, pp. 10–29.

PERLMAN, JACOB, "Some Problems in the Construction of an Index Number of Rents," *Journal of the American Statistical Association*, 29 (185A), March, 1934, pp. 24–29.

PERRY, E. G., and SILVERMAN, A. G., "A New Index of the Physical Volume of Canadian Business," *Journal of the American Statistical Association*, 24 (166), June, 1929, pp. 127–143.

SNYDER, CARL, and PISER, LEROY M., "The Index of the Volume of Trade: Third Revision," *Journal of the American Statistical Association*, 26 (176), December, 1931, pp. 436–442.

——"Outline of Index Numbers of Wholesale Prices in Foreign Countries," *Monthly Labor Review*, 31 (4), October, 1930, pp. 42–58.

WALD, A., "A New Formula for the Index of Cost of Living," *Econometrica*, 7 (4), October, 1939, pp. 319–331.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Making use of the data of Exercise 1, Chapter VIII, page 200, compute index numbers of value, quantity, and price by Fisher's method.

2. Making use of the same data as above, compute index numbers of quantity and price by the relative method, base-weighted, using both the arithmetic and the geometric mean.

3. Making use of the data of Exercise 2, page 200, compute index numbers of quantity, and price by Fisher's method.

4. Making use of the data of Exercise 3, page 200, compute quantity and price indexes by Fisher's method (1890 = 100).

ANSWERS TO EXERCISES

1.

YEAR	V	Q_i	P_i
1933	100	100.0	100.0
1934	150	131.8	113.8
1935	268	119.4	224.5
1936	384	146.4	262.4
1937	200	100.0	200.0

2.

	Arithmetic		Geometric	
YEAR	Q	P	Q	P
1933	100	100	100.0	100.0
1934	132	114	130.1	108.3
1935	116	218	110.6	201.3
1936	144	258	142.0	239.6
1937	100	200	100.0	170.5

3. (a)

YEAR	V	Q_i	P_i
1926	100	100.0	100.0
1927	156	188.8	82.6
1928	164	191.2	85.8
1929	200	151.6	132.1

(b)

YEAR	V	Q_i	P_i
1926	100	100.0	100.0
1927	210	145.0	145.0
1928	360	177.3	203.4
1929	250	88.1	283.9

(c)

YEAR	V	Q_i	P_i
1926	100	100.0	100.0
1927	170	161.0	105.6
1928	180	144.2	125.0
1929	200	200.0	100.0

(d)

YEAR	V	Q_i	P_i
1935	100	100.0	100.0
1936	150	117.7	127.5
1937	145	96.8	150.0
1938	100	115.4	90.7

4.

YEAR	V	Q	P	YEAR	V	Q	P
1890	100	100.0	100.0	1903	164	148.5	110.4
1891	114	114.0	100.0	1904	188	157.3	119.5
1892	90	100.0	90.0	1905	194	176.2	110.1
1893	62	77.7	79.8	1906	224	195.3	114.8
1894	60	83.7	71.7	1907	212	180.8	117.2
1895	78	104.7	74.5	1908	182	166.7	109.2
1896	76	117.4	64.8	1909	216	184.5	117.0
1897	86	123.4	69.7	1910	234	193.9	120.7
1898	96	129.2	74.4	1911	216	190.4	113.5
1899	110	122.1	90.1	1912	280	211.0	132.7
1900	134	135.0	99.3	1913	276	208.6	132.4
1901	110	122.1	90.1	1914	170	186.5	91.2
1902	178	161.9	110.0				

B. PROBLEMS

5. The table below summarizes facts with respect to the receipts and prices of the three major cereals at 13 principal grain markets in the United States, during a recent period.

Year	Grain receipts in millions of bushels			Grain prices in dollars		
	Wheat	Corn	Oats	Wheat	Corn	Oats
1928	518.7	289.4	137.4	1.18	0.92	0.44
1929	418.8	254.3	134.5	1.33	0.83	0.44
1930	483.7	194.9	99.5	0.83	0.60	0.35
1931	360.2	146.2	67.4	0.68	0.36	0.22
1932	270.8	237.5	100.6	0.60	0.35	0.22
1933	201.4	210.6	64.1	0.94	0.52	0.36

(*a*) Prepare simple relatives expressing changes in prices, quantities, and values of each of the grains in this period, using 1928 as the base.

(*b*) Using the method of weighted relatives, prepare an index showing the composite changes in grain prices. Use 1928 values as weights.

(*c*) Using the common aggregative method, prepare an index of composite grain prices, base-weighted, using 1928 as the base.

(*d*) In a similar manner, prepare an index of quantities, base-weighted as to price, using 1928 as the base.

(*e*) Prepare an index of composite prices using Fisher's ideal method.

6. Data summarized below represent relative prices of farm products received by farmers in the state of Iowa for the periods indicated, together with income weights applicable to years and to specified months.

(*a*) Check the price relatives for each commodity for the 5-year and annual periods indicated by reference to Problem 5, page 203.

(*b*) Average these relatives by use of geometric means employing annual value weights, as given below; 1910–1914 is taken as the base.

(*c*) Similarly calculate composite price indexes for appended monthly data, using appropriate weights.

(*d*) Compare the price indexes obtained for the years 1925–1938 with those obtained by the aggregative method in Problem 5, page 203. How might these index numbers theoretically be expected to compare?

RELATIVE FARM PRODUCTS PRICES, IOWA, 1910–1939
(1910–1914 = 100)

Year	Hogs	Cattle	Sheep	Corn	Oats	Wheat	Hay	Butter	Eggs	Poultry
1910–14	100	100	100	100	100	100	100	100	100	100
1915–19	172	151	178	203	157	196	139	148	164	161
1920–24	119	120	134	136	117	146	129	161	155	180
1925	151	133	165	170	110	166	114	161	154	184
1926	159	125	153	115	99	151	142	165	154	201
1927	131	140	145	140	119	143	136	173	124	184
1928	117	171	154	153	122	127	123	181	148	204
1929	129	169	145	148	113	126	116	181	154	198
1930	120	144	104	132	96	91	95	142	112	157 ,
1931	77	102	64	81	61	52	84	106	83	140
1932	44	77	44	44	44	44	77	79	65	97
1933	46	68	50	51	64	81	55	83	65	77
1934	57	78	62	108	113	103	114	94	77	105
1935	119	128	84	136	93	103	111	114	122	146
1936	127	115	82	144	99	121	. 90	130	107	140
1937	131	142	90	170	99	121	102	134	106	160
1938	105	122	72	78	58	72	67	110	90	134
1939										
Mar.	97	133	83	66	67	66	54	94	81	129
June	81	128	73	76	78	74	53	94	69	120
Sept.	97	138	78	91	84	85	56	106	86	129
Dec.	. 66	131	90	83	96	100	60	118	85	100

WEIGHTS FOR FARM PRODUCTS PRICES, BASED ON INCOME, 1925–1929

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Year
Hogs	54.0	52.7	47.0	41.6	39.7	41.5	40.9	35.7	36.3	37.2	47.3	50.8	44.3
Cattle	16.8	16.3	18.7	19.9	19.9	18.8	18.0	17.2	19.1	18.3	17.5	17.0	18.0
Sheep	1.0	0.7	0.4	0.3	0.3	0.4	0.5	0.8	0.9	1.1	1.3	1.2	0.8
Corn	11.1	11.9	8.9	6.7	7.3	9.1	10.1	10.2	13.2	13.7	9.4	11.7	10.3
Oats	2.2	2.3	2.7	2.7	2.4	2.1	3.3	10.0	5.3	4.9	2.2	2.3	3.5
Wheat	0.5	0.6	0.5	0.4	0.4	0.7	3.0	3.0	2.8	1.8	1.0	0.7	1.2
Hay	0.5	0.6	0.7	0.7	0.5	0.5	0.8	0.9	0.8	0.9	0.7	0.6	0.7
Butter	9.0	10.1	12.0	13.0	15.1	15.7	14.3	14.3	13.1	12.4	10.1	7.7	12.0
Eggs	1.4	2.6	7.5	13.5	12.9	8.6	6.3	5.1	4.1	2.6	1.0	0.7	5.3
Poultry	3.5	2.2	1.6	1.2	1.5	2.6	2.8	2.8	4.4	7.1	9.5	7.3	3.9

7. Utilizing the data of Problem 11, page 206, calculate value weights as consumption times base prices, and recompute the price index for February, 1939, by the relative method, geometric mean. Which answer do you consider most accurate? Why?

CHAPTER X

ELEMENTARY TRENDS

In preceding chapters, attention has béen given to the problem of averaging, by which the common characteristics of a group are represented by a single magnitude. In the present chapter, the concept of averaging is extended, and the problem involves the discovery of a representative line or curve, rather than a single point or item. Such a curve is intended to represent the *average or composite direction of change* in a series of data extending over a period of time. In later chapters, a similar procedure will be adapted to data other than time series. In all such cases, a line or curve thus employed is known as a *trend*.

Uses of trend analysis.—The possible utility of a trend in connection with the data of business may be suggested by a concrete example. Consider, for instance, a canning concern which makes use of large quantities of sugar. Such an organization would·be interested in the trend of sugar production in the United States over a period of years, inasmuch as that trend would throw some light on probable future production.

If, in 1930, the production of sugar in the United States had been plotted at fairly regular intervals from 1874 to 1930, and if a trend line indicating the composite direction of change among these points ·had been sketched freehand, the resulting chart would have resembled Fig. 10·1. The trend might have been projected through the next few years, as a preliminary estimate of production in the immediate future. If the current reports concerning growing crops in each succeeding season were considered, the canning concern might find such a trend most useful as a tentative base, or statistical *normal*, from which to develop fairly accurate annual estimates.

Another manner in which trends are frequently utilized may be illustrated by reference to the composite index of industrial

production which has been charted in Fig. 10·2. Since comparable data in this series are available back to 1900, it is possible to calculate a trend that may reasonably be extrapolated
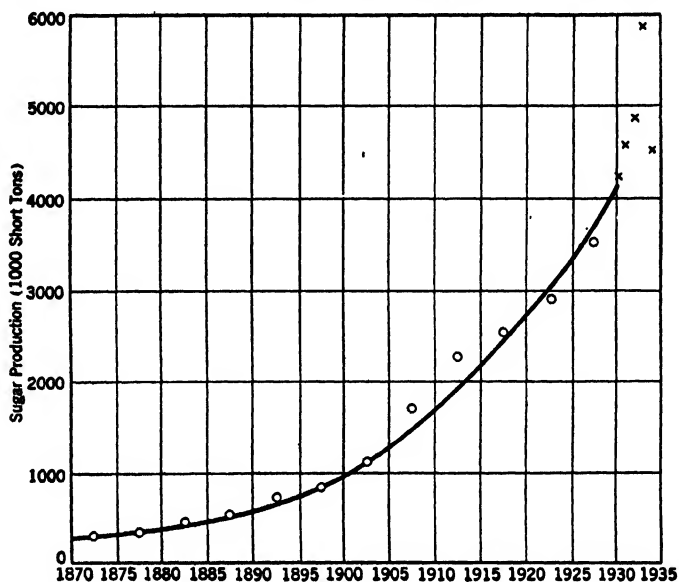

FIG. 10·1.—Sugar Production, United States and Outlying Territory, 5-Year Annual Averages, 1870–1929, and Annual Averages, 1930–1934, with Freehand Trend, 1870–1930. (Data from *Statistical Abstract of the United States*, 1935, p. 638.)
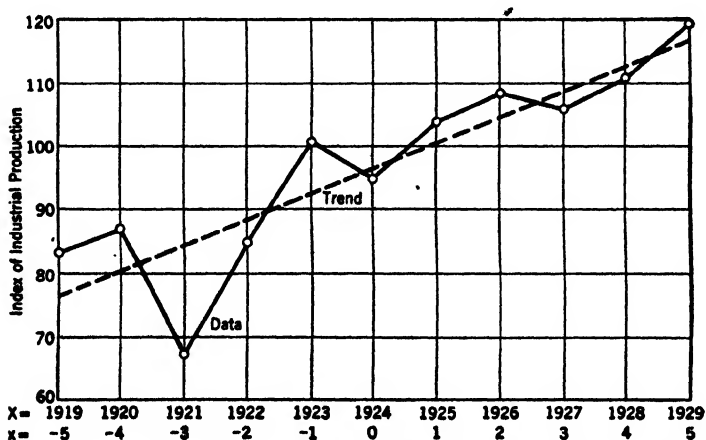

FIG. 10·2.—Straight-Line Trend of Industrial Production, United States, 1919–1929. (Data: Federal Reserve Board Index, without 1940 revisions.)

a few years in advance.[1] Actual measures of industrial production, as they appeared, could be compared with such a trend and expressed as percentages of *normal*. If, for example, such a trend based on more extensive data is projected to 1935, the estimated normal for that year is 127. The actual figure for that year is 90, so that it may be said that production in that year represented 71 per cent of normal, as the latter is estimated from the trend of earlier years. Such a projected trend should, of course, be based upon adequate data, representing a number of years, and it should always be checked with any additional pertinent facts.[2]

Two of the most common uses of trends will be clear from these illustrations. Trends are used to project what appears to be the composite direction of change into the future. They are also used as a base which averages out minor irregularities to make the general direction of change stand out and with which given items representing specific periods may be compared and described as percentages of what is assumed to be normal. When the data utilized in defining the trend cover a considerable period of time, seldom less than a decade, the measure of composite change is commonly called a *secular* trend. In statistical analysis, the secular trend is most frequently used (1) to provide a preliminary estimate of future items, and (2) as a basis from which seasonal and larger cyclical fluctuations may be measured. In addition to these uses, trends and the methodology of trend analysis serve a variety of purposes in more complicated statistical analysis, particularly in connection with the procedures of correlation, which are described in later chapters.

Types of trends.—The statistical problem of trend fitting involves the discovery and perfection of methods for fitting lines or curves that accurately and effectively picture the general

---

[1] In 1940, this series was revised for the period beginning in 1923, and the data of the revised series are summarized in Table 8·3, page 187. Because the revision does not extend back beyond 1923, it is questionable whether it should be used as a basis for trend analysis.

[2] The comparatively short series here employed approximates the long-time trend of American business as calculated by Dr. Carl Snyder ("Capital Supply and National Well Being," *American Economic Review*, June, 1936, pp. 195–224).

direction of change of the specific data being analyzed. Given
the data, the first questions are: What shape of curve most
effectively represents the course they take? Is a straight line
the most accurate and effective figurative representation? Or is
one of the many possible curved lines required to portray realisti-
cally the changes? When these questions have been answered,
a second follows directly. It asks how the selected curve may
be fitted to the particular data under consideration.

The answers which statistics makes to these questions are
described most satisfactorily by reference to the meaning of
trend lines. The trend always portrays a relationship between
two variables. In the cases now under consideration, one of
these is time, and it is desirable to illustrate the manner in
which the other variable changes, or has changed, with the pas-
sage of time. It may, for instance, be desired to describe the
general trend in the varying volume of industrial production in
the United States, or of sugar production, as suggested in earlier
examples. Again, it may be that the trend is required to show
changing levels of money wages, or of living costs, or historic
tendencies in imports or exports of various goods. In all such
cases, one variable is time; the other is represented by the actual
figures or indexes for weeks, months, quarters, or years. If the
relationship is charted, the time intervals (the so-called inde-
pendent variable) are ordinarily measured along the base line or
X axis, while values of the business data involved (generally
described as dependent) are recorded above the appropriate
dates and measured by a scale on the Y axis.

A trend fitted to the Y items describes the manner in which,
in general, they appear to vary with time. Because there are
so many possible relationships among various series of data, the
possible shapes of trend lines are infinite in number, but statis-
tical analysis generally makes use of only a few fairly simple
types of curves.

In some cases, illustrated by the long-time growth of popu-
lation under reasonably stable conditions and by the accumu-
lation of savings at compound interest rates, the essential func-
tional relationship between the variables actually dictates the
shape of the trend, so that its appropriate form (not, of course,

its specific dimensions) is known in advance, without reference
to the particular data under consideration. Generally, however,
no such established relationship between the variables prevails,
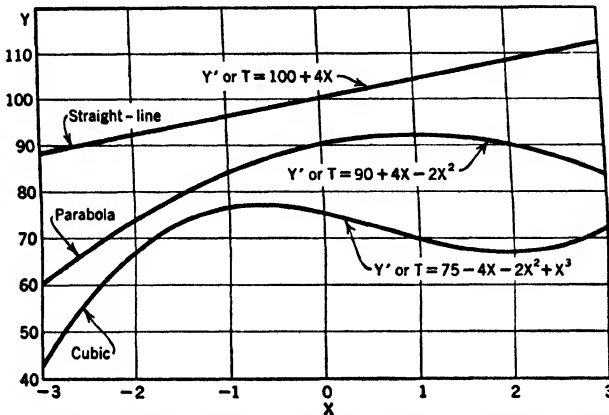
Fig. 10·3a.—Three Typical Trend Curves of the Potential Series.

and it becomes necessary to select an appropriate trend largely
upon the basis of inspection of the data involved. The range
of possible curves that may be used to represent such data is

Fig. 10·3b.—Three Typical Trend Curves of the Exponential Series.

suggested in Figs. 10·3a and 10·3b, where several of the most
usable types of curves are illustrated. The problem of deciding
which curve is to be used is the first problem of trend analysis.

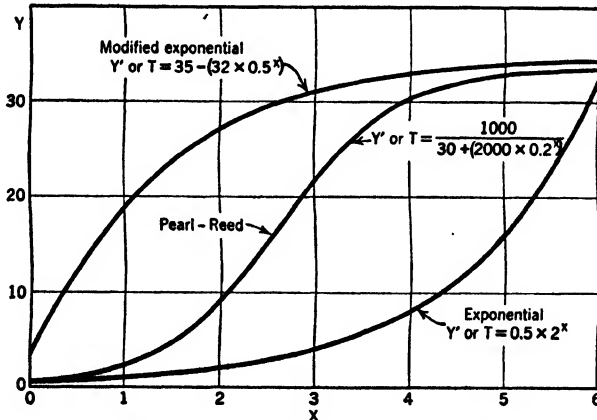When the type of trend that appears most appropriate has been determined, it is necessary to fit that type to the specific data under consideration. There are many methods of effecting such a fitting, ranging from simple, freehand sketching to rather involved mathematical procedures. In practically every case, the logical first step is the preparation of a chart upon which items for each time period are clearly indicated. Such a chart, in effect, states the problem by indicating how complicated the trend must be in order fairly to represent the composite change featuring the variables.

The most commonly used methods of fitting various types of trends are described in this and the next chapter. Throughout the discussion, it will be well to keep clearly in mind the fundamental objective in all trend fitting, i.e., the accurate, faithful description of the covariation featuring the two variables under consideration.

`Freehand trend fitting.—The simplest method of fitting trends is that which involves merely drawing them in such a manner as to make them appear to be as fairly representative of the original data as possible. The method has been illustrated in Fig. 10·1. The data are first plotted, indicating each item with a dot, cross, or small circle. When their general pattern has been observed, the trend line is sketched in, major consideration being given to the problem of so placing the line (straight or curved) as to represent best the changes that have taken place in the data in the period under consideration.

When the trend has been sketched, it may be smoothed and defined more clearly by means of tools generally available to the draftsman, particularly flexible or curved guides and rules.

A freehand trend has certain advantages. It is fitted simply and quickly, and, if care is used, it may provide an excellent representation of the composite change in the series. It does not depend upon complicated mathematical procedure or involved assumptions as to its propriety. In effect, it says, "this is what the trend appears to be," and it expresses no reasons for its direction or its continuance. For these reasons, the freehand trend is sometimes preferred to the more complicated mathematical types to be described in subsequent portions of this dis-

cussion. Its chief limitation is the fact that it may lead to disputable or biased results.

The moving average.—A flexible trend which may readily be adjusted to consecutive items of a time series is called a moving average. Each item in this trend is the mean of a certain number of consecutive items of the data. It is placed at the middle of the time interval they cover. For example, a three-term moving average of the annual series, *A, B, C, D, E*, etc., is
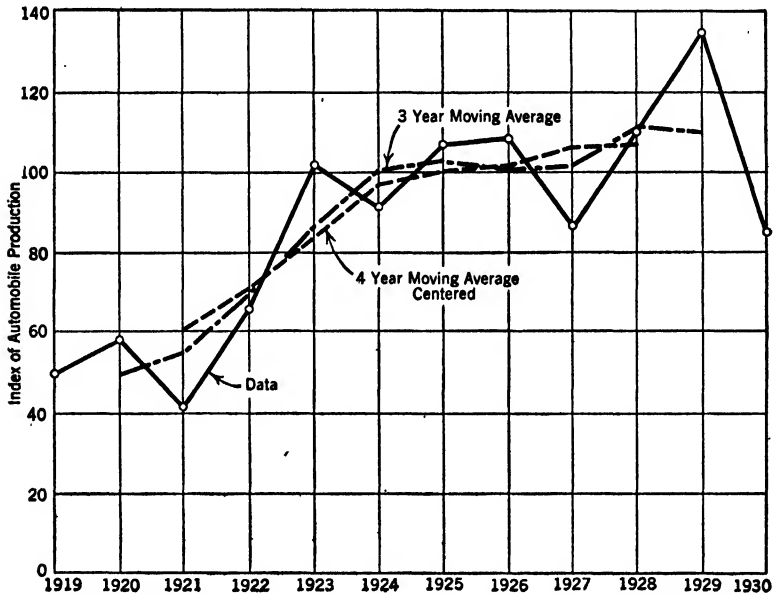
Fig. 10·4.—Three-Year and Four-Year Centered Moving Averages. Data: Index numbers of automobile production, adapted from the annual report of the General Motors Corporation, 1936.

begun by averaging *A, B,* and *C,* and regarding this average as the trend item in the second year. Similarly, *B, C,* and *D* are averaged to obtain the trend item in the third year, and *C, D,* and *E* are averaged to obtain the trend item in the fourth year, and so on to the end of the series. It will be seen that this procedure fails to provide trend items for the first and last years. A five-term annual moving average would obviously begin the trend at the third year and would fail to provide trend figures for the first two and the last two years.

If the moving average covers an even number of items, it naturally centers between two points on the time scale. One way of centering such an average involves taking one more item and giving only half-weight to the first and last items. For example, a four-term moving average of the annual data A, B, C, D, E, F, etc., begins with the weighted average $(A + 2B + 2C + 2D + E) \div 8$, as the trend item of the third year. The weighted average $(B + 2C + 2D + 2E + F) \div 8$ is the trend item of the fourth year, etc. Another and more commonly used device arbitrarily centers each average of an even number of items on one of the two center items. This method is satisfactory and justifiable if the error thus occasioned is relatively insignificant. The calculation of three- and four-term moving averages is illustrated in Example 10·1 (see Fig. 10·4).

The moving average is not commonly utilized where the trend is desired as a basis of comparison with current data, or where it is to be extrapolated, because it fails to provide the necessary items. It has, however, considerable use as a trend in connection with the calculation of indexes of seasonal variation, where monthly data are compared with the moving average. This use of the moving average will be given additional attention in the next chapter. For this purpose, each average covers one year, and it is the average of monthly, quarterly, or other seasonal data. The moving average is also useful in describing a trend from which cyclical variations are to be eliminated, in which event the number of terms in each average is generally chosen to coincide as nearly as possible with the length of cyclical swings in the data. For example, if annual data seem to reflect a five-year cycle, then a five-term moving average (including five annual items) would tend to effect a smoothed trend as far as this cycle is concerned. Similarly, a cyclic tendency of any number of years would be smoothed by the use of a moving average covering a corresponding number of annual items.

Trend equations.—The problem of securing an effective adjustment of a trend to specific data is solved, in a large part of all statistical procedure, by means of mathematical equations. Such equations may appear formidable to the uniniti-

EXAMPLE 10·1

THE MOVING AVERAGE

Data: Index of automobile production, United States, 1919–1930 [1]

Years	Index of automobile production	3-year totals	3-year moving average	4-year totals multiplied by 2; 5-term wts.: 1, 2, 2, 2, 1	4-year moving average centered
1919	50	. . .	. . . . .	. . .	. . . . .
1920	58	149	49.7	. . .	. . . . .
1921	41	165	55.0	482	60.2
1922	66	209	69.7	567	70.9
1923	102	259	86.3	666	83.2
1924	91	300	100.0	774	96.8
1925	107	306	102.0	800	100.0
1926	108	301	100.3	803	100.4
1927	86	304	101.3	850	106.2
1928	110	331	110.3	855	106.9
1929	135	330	110.0	. . .	. . . . .
1930	85	. . .	. . . . .	. . .	. . . . .

Given a regular time series, A, B, C, D, etc., the three-term moving average is:

Corresponding to A, none

Corresponding to B, $(A + B + C) \div 3 = (50 + 58 + 41) \div 3 = 49.7$

Corresponding to C, $(B + C + D) \div 3 = (58 + 41 + 66) \div 3 = 55.0$
  etc.

The four-term centered moving average is:

Corresponding to A, none

Corresponding to B, none

Corresponding to C, $(A + 2B + 2C + 2D + E) \div 8$
$$= [50 + (2 \times 58) + (2 \times 41) + (2 \times 66) + 102] \div 8 = 60.2$$
  etc.

ated, but they are nothing more than mathematical descriptions of the various types of lines, straight or curved, which represent trends. The simplest and most common is the equation for a straight line, which appears as:

$$T = a + bX$$

[1] 1932 Annual Supplement, *Survey of Current Business*, United States Department of Commerce, pp. 8 and 9.

The T represents the trend value. The a is a value that defines the height of the line at the date selected as the *origin* or starting point, and the b is the measure of the slope that is introduced with each additional time interval, X. Thus, the value of the trend at any given time is the height at the point of origin, a, plus the product of the measure of slope, b, and the number of time periods intervening between the given date and that of the origin.

In a similar manner, more complicated curved trends may also be represented by equations descriptive of the particular types of curves required. Several of these more complex equations will be described in the next chapter.

The problem in fitting the curves to specific data consists of discovering appropriate values for the constants, the a, b, etc., in these equations. Since X values are given, once the values of the constants are known the trend is readily defined by substituting these values in the trend equation.

Time centering.—The process of time centering consists of expressing time (X) in terms of deviations from a central date in the series. The deviations are then designated as x to distinguish them from the uncentered time items. For example, the series of years (X): 1924, 1925, 1926, 1927, and 1928 is centered at $1926 = 0$ (the average or central item), and the individual years (x) are equal to -2, -1, 0, 1, and 2, respectively. If the problem involves an even number of years, say 1924, 1925, 1926, and 1927, of which there is no central item, the average might be calculated as 1925.5, and the individual items expressed as -1.5, -0.5, 0.5, and 1.5. In any case, centering or at least rescaling X materially reduces the amount of calculation involved in discovering trend values, and the device is useful for almost all types of trend fitting.

THE STRAIGHT-LINE TREND

A discussion of mathematical methods of trend fitting logically begins with a consideration of the straight-line trend, since this type presents the simplest problems. There are several methods of fitting linear trends, but the most common are those of selected points, semi-averages, and least squares.

The selected-points method.—Before considering methods requiring time-centering and a strictly mathematical equation, attention may be called to a method employing a very simple equation. As has already been suggested, a trend may be approximated by drawing, freehand, an appropriate line or curve through the charted data. In the case of a straight-line trend, such an estimate is improved by a procedure known as the method of *selected points*, which may be described, as it applies to straight-line trends, as follows: The values of two points, P_1 and P_2, located near the beginning and the end of the period under consideration, respectively, are noted on the Y scale of the chart. These points are estimated to lie on the appropriate trend. Then the difference between the Y values of these two points, readily noted from the chart, is divided by the number of years (t) or other time period separating them to secure a measure of the *slope* of the line, which is generally designated as b. That is,

$$b = \frac{P_2 - P_1}{t}$$

Trend values for each time item are obtained by beginning with P_1 and adding b for each successive time period, and subtracting b for each preceding time period. In other words, these values are obtained as

$$T = P_1 + b(X - X_0)$$

where X indicates any given year or time period, and X_0 represents the time item at which P_1 is located.

The method may be illustrated by reference to Fig. 10·2, if the trend as drawn there is ignored. The first point, P_1, might be selected as representing 1920 and estimated as 80, and P_2 as representing 1928 and estimated to be 112. Since $t = 1928 - 1920 = 8$, the slope is calculated as

$$b = \frac{P_2 - P_1}{t} = \frac{112 - 80}{8} = 4$$

Using this figure, the trend may be calculated for any year. For instance, that for 1929 is

$$T_{1929} = P_1 + b(X - X_0) = 80 + 4(1929 - 1920) = 116$$

In the same manner, the trend values for the other years of the period are determined as

Years:	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929
Trend:	76	80	84	88	92	96	100	104	108	112	116

To test this estimate of the trend, the average of the original data ($M_Y = \Sigma Y \div N$) is compared with the average of the trend items ($M_T = \Sigma T \div N$). They should be practically the same. If they differ materially, the quantity $M_Y - M_T$ may be added algebraically to each trend item. It will be noted that b may be either negative or positive, for the slope may be either downward or upward to the right.

The method of selected points, even though the correction suggested is made, is no more than an estimate, and more precise ways of trend fitting are frequently required. However, its use is not limited to the straight line; it is also applicable to more complex trends.

Semi-averages method.—As has been indicated, the more exact fitting of a straight-line trend requires discovery of the values of the constants, a and b. When time is centered, that is, when the origin is taken at the mid-point of time, a necessarily becomes the average of the Y data. That is, on an x scale (but not on an X scale),

$$a = M_Y = \frac{\Sigma Y}{N}$$

If the origin is not at the central point of time, the equation for a will require adjustment, as will be pointed out later.

The value designated as b is the slope of the line, or the number of points it rises or falls in one time unit. A fairly simple method of determining this slope is that provided by the method of semi-averages. Graphically, this method simply averages separately the first and second halves of the Y data (subtotals Σ_1 and Σ_2) and draws a line through these "semi-averages." But in practice, on account of complications arising when N is odd, it is better to utilize the simple formula

$$b = \frac{\Sigma_2 - \Sigma_1}{m(N - m)}$$

where Σ_1 and Σ_2 are the respective sums of Y items before and after the time center (the central item is omitted if N is odd), and m is the number of items included in each subtotal.

The method of semi-averages may be conveniently illustrated by reference to the data of part one of Example 10·2, page 237. The data used in the example, assumed for purposes of illustration, are sufficiently simple so that the procedure may be clearly seen. The value of a may be readily discovered as follows:

$$a = M_Y = \frac{\Sigma Y}{N} = \frac{65}{5} = 13$$

and,

$$b = \frac{\Sigma_2 - \Sigma_1}{m(N - m)} = \frac{30 - 22}{2 \times 3} = 1.3\underline{3}$$

The trend equation for these data is, therefore,

$$T = 13 + 1.3\underline{3}x$$

which definitely describes the trend as having a height of 13 at the central date, 1934, and rising 1.3$\underline{3}$ points [1] from any one year to the next. The trend for any given year may, therefore, be readily found by substituting the date (expressed in terms of deviations from the central date) in the trend equation. Thus, in 1936, when $x = 2$, the trend is defined by the equation

$$T = 13 + (1.3\underline{3} \times 2) = 15.6\underline{6}$$

This result is not quite the same as that obtained by the method of least squares described later. It differs from the least-squares trend because of an implicit change in the weighting of the Y data.

It may be noted that the method of semi-averages is an elementary example of what is sometimes called the method of grouped data, in which items are divided into two or more groups, according to the type of trend to be fitted. For those who are interested in this type of procedure, additional atten-

---

[1] Underscoring indicates repetition of the designated digits. For example, 1.$\underline{3}$ means that if calculations were carried to additional decimal places, the *three* would be repeated and the result would be 1.33333. ... Similarly, a figure which appears as 1.8142857142857...may more conveniently be written 1.8$\underline{142857}$.

tion is given to the method of grouped data in the Appendix, page 530.

Method of least squares.—The methods of determining trends discussed so far cannot be regarded as mathematically precise, because they do not weight each item in a time series in accordance with its position. They simply estimate comparative levels, or lump the items into groups. A more justifiable procedure, from the mathematical standpoint, is the method of least squares. Proof of the efficiency of this method is available in the logic of its derivation, but it is sufficient here to note that the method so fits the trend that the standard deviation of the differences between trend items and actual items for the same dates, or $\Sigma(Y - T)^2$, is a minimum. That is, as measured by the squared deviations, the least-squares straight-line trend is graphically closer to the data than any other straight line that might be drawn.

The method of least squares is distinguished from that last described principally by the procedure used in finding the measure of slope. Time may be centered, and a found as the mean of the Y series. The value of b, however, is discovered through the use of what are described as normal equations, so prepared that their solution assures the closeness of fit mentioned in the preceding paragraph. For the straight-line trend, these equations are:

$$Na + b\Sigma X = \Sigma Y$$

$$a\Sigma X + b\Sigma X^2 = \Sigma XY$$

As will be explained later, these equations may be solved, without time centering, by substituting the required values derived from the data. But with time centering, they may be reduced to

$$Na = \Sigma Y$$

$$b\Sigma x^2 = \Sigma xY$$

or, for greater convenience,

$$a = \frac{\Sigma Y}{N} = M_Y$$

$$b = \frac{\Sigma xY}{\Sigma x^2}$$

<div style="text-align:center">EXAMPLE 10·2</div>

THE LEAST-SQUARES STRAIGHT-LINE TREND

I. Data (assumed for purposes of illustration): Number of employees in a small firm as of July 1, each year.

Year	Number of employees	Years centered	Summations		Trend equation and trend
X	Y	x	x^2	xY	$a + bx = T$
1932	10	-2	4	-20	$13 - 2.8 = 10.2$
1933	12	-1	1	-12	$13 - 1.4 = 11.6$
1934	13 .	0	0	0	$13 \quad .0 = 13.0$
1935	14	1	1	14	$13 \quad 1.4 = 14.4$
1936	16	2	4	32	$13 \quad 2.8 = 15.8$
$M = 1934$	65		10	14	

$$a = \frac{\Sigma Y}{N} = \frac{65}{5} = 13.0 \qquad b = \frac{\Sigma xY}{\Sigma x^2} = \frac{14}{10} = 1.4$$

II. Data: Wholesale prices in the United States (1926 taken as 100)

Year	Prices	Years centered	Summations		Trend equation, trend
X	Y	x	x^2	xY	$a + bx = T$
1923	100.6	-2.5	6.25	-251.50	$99.05 + 2.22 = 101.27$
1924	98.1	-1.5	2.25	-147.15	$99.05 + 1.33 = 100.38$
1925	103.5	-0.5	0.25	$- 51.75$	$99.05 + 0.44 = 99.49$
1926	100.0	0.5	0.25	50.00	$99.05 - 0.44 = 98.61$
1927	95.4	1.5·	2.25	143.10	$99.05 - 1.33 = 97.72$
1928	96.7	2.5	6.25	241.75	$99.05 - 2.22 = 96.83$
$M = 1925.5$	594.3	0	17.50	-15.55	594.30

$$a = \frac{\Sigma Y}{N} = \frac{594.3}{6} = 99.05 \qquad b = \frac{\Sigma xY}{\Sigma x^2} = -\frac{15.55}{17.50} = -0.8886.$$

III. Data: As in II above; time unit taken as half years

Year	Prices	Half years centered	Summations		Trend equation, trend
X	Y	x	x^2	xY	$a + bx = T$
1923	100.6	-5	25	-503.0	$99.05 + 2.22 = 101.27$
1924	98.1	-3	9	-294.3	$99.05 + 1.33 = 100.38$
1925	103.5	-1	1	-103.5	$99.05 + 0.44 = 99.49$
1926	100.0	1	1	100.0	$99.05 - 0.44 = 98.61$
1927	95.4	3	9	286.2	$99.05 - 1.33 = 97.72$
1928	96.7	5	25	483.5	$99.05 - 2.22 = 96.83$
$M = 1925.5$	594.3	0	70	$- 31.1$	594.30

$$a = \frac{\Sigma Y}{N} = \frac{594.3}{6} = 99.05 \qquad b = \frac{\Sigma xY}{\Sigma x^2} = \frac{-31.1}{70} = -0.4443 \text{ (per half-year)}$$

The process of fitting the trend by the method of least squares necessarily begins with the discovery of the values required for the solution of these equations, i.e., ΣY, ΣxY, and Σx^2. A convenient procedure for this purpose is outlined in Example 10·2, and the trends calculated in the example are plotted in Figs. 10·5 and 10·6. In order to make the procedure clear, very brief and simple series have been used in these illustrations. In practice, of course, such brief series could not be regarded as satisfactory for trend analyses.

In part I of Example 10·2, the data extend through an odd number of consecutive years. The time scale, therefore, is readily expressed in deviations about the central year, 1934. It is more accurate, however, to define this central point as the average of the X's, since the central point might not occupy the central position in the column if the X items were irregular. In general, the x column expresses the deviations of each X from the average X; that is, $x = X - M_x$.[1]

In part II of the same example the number of years is even, hence the mean of the X column is fractional. Otherwise, the calculation is as before. In this illustration, b is negative, and the trend is, therefore, downward. Occasionally it happens that $b = 0$, in which case the trend is horizontal at the a level.

In part III the same trend is again fitted, but in such a way as to avoid fractions in the x column. This is accomplished by the simple expedient of taking the time unit as a half year. As a result each x is doubled, but b is half what it was before, since it expresses the change in the trend line during half as long an interval. The values of bx, however, are not changed, and the resulting trend is identical with that previously discovered.

The device of changing the time unit may be applied under other circumstances. For example, data may be expressed in index numbers or aggregates for five-year, ten-year, or other periods. In such cases the time unit may be a year, a two-and-a-half-year interval, or any other period that is convenient. As

---

[1] It should be noted that, in calculating $b = \Sigma xY/\Sigma x^2$, the decimals should be carried beyond the accuracy required in the trend, to minimize the error in the bx column. It should also be noted that Σxy (both X and Y centered) equals ΣxY (only X centered).
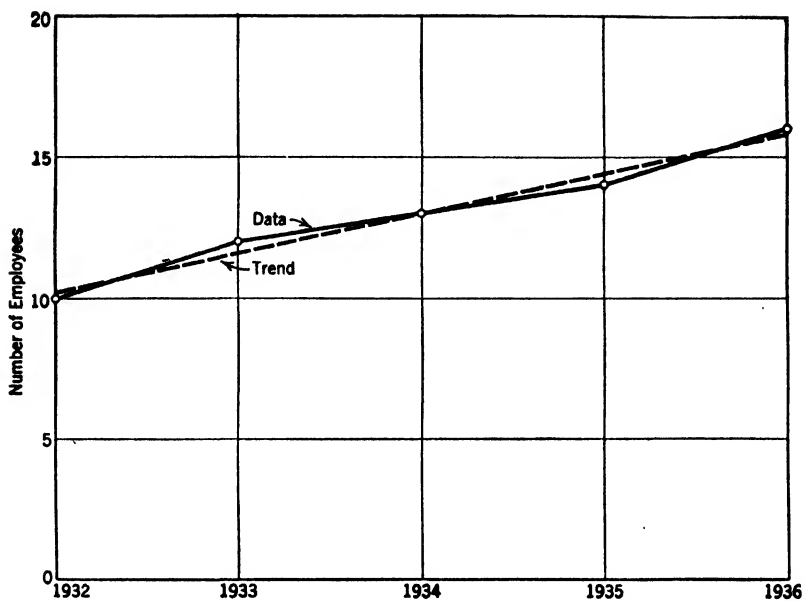
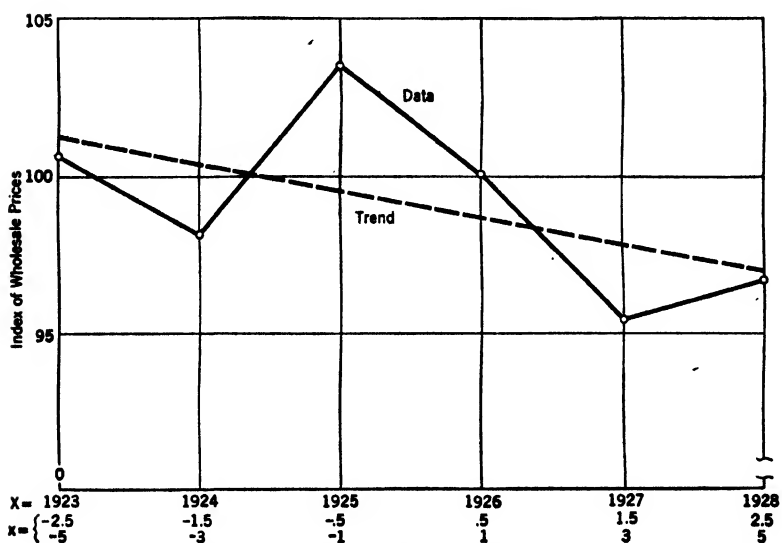Fig. 10·5.—Straight-Line Trend Fitted to Data of Example 10·2, Part I.

Fig. 10·6.—Straight-Line Trend Fitted to Data of Example 10·2, Part II.

in the illustration, the value of b expresses the slope of the trend for the selected interval, and the trend itself is unchanged by the change of the time scale.

Interpolating and extrapolating.—When the trend equation has been determined, whether by the least squares or the semi-averages method, it is possible to interpolate or extrapolate a trend item at any required point on the time scale merely by determining the value of x at that point and substituting it in the equation. Thus in Example 10·2, part I, the trend equation is

$$T = 13 + 1.4x \text{ (origin 1934)}$$

If the trend at the beginning of 1933 ($x = 1932.5 - 1934 = -1.5$) is required, it may be found as

$$\text{(Jan. 1, 1933)} T = 13 + 1.4(-1.5) = 13 - 2.1 = 10.9$$

If the assumption could be made that the same general trend would continue, the items for future dates could be readily estimated by projecting the line to the appropriate ordinates. For 1937 ($x = 3$) the calculation would be as follows:

$$\text{(July 1, 1937)} T = 13 + 1.4(3) = 13 + 4.2 = 17.2$$

Building up the straight-line trend.—As illustrated in Example 10·2, the items of the straight-line trend were computed by multiplying the x column by b and adding the products to a, thus solving the trend equations for the required successive values of x. In practice, however, it is generally more convenient to compute the trend items by beginning at the origin, where the trend obviously is a, and, if N is odd, adding b successively forward to the last required item and subtracting it successively to the earliest required item. This addition or subtraction may be accomplished rapidly and accurately on a calculator by obtaining the value of b to several decimal places, and rounding the results to the number of decimals required. If N is even, as in Example 10·2, part II, one-half b must first be added and subtracted in order to obtain the trend items between which the origin falls, after which b is added forward

EXAMPLE 10·3

STRAIGHT-LINE TREND

GENERAL LEAST-SQUARES SOLUTIONS

Data: Index of production, simplified for illustrative purposes

I. Solution by normal equations:

YEAR	Y	X	X^2	XY	$a + bX = T$
1923	100	0	0	0	$98 + 4(0) = 98$
1924	96	1	1	96	$98 + 4(1) = 102$
1925	108	2	4	216	$98 + 4(2) = 106$
1926	116	3	9	348	$98 + 4(3) = 110$
1927	110	4	16	440	$98 + 4(4) = 114$
1928	118	5	25	590	$98 + 4(5) = 118$
	648	15	55	1,690	648

Normal equations

$$Na + b\Sigma X = \Sigma Y = 6a + 15b = 648 \tag{1}$$

$$a\Sigma X + b\Sigma X^2 = \Sigma XY = 15a + 55b = 1,690 \tag{2}$$

Dividing (1) by 6, (2) by 15, subtracting and solving

$$a + 2.50b = 108$$
$$a + 3.66\underline{6}b = 112.66\underline{6}$$
$$\overline{}$$
$$1.16\underline{6}b = 4.66\underline{6}$$
$$b = 4$$
$$a + 2.50 \times 4 = 108$$
$$a + 10 = 108$$
$$a = 108 - 10$$
$$a = 98$$

II. Solution by centering equations:

YEAR	Y	X	X^2	XY	T
1923	100	0	0	0	98
1924	96	1	1	96	102
1925	108	2	4	216	106
1926	116	3	9	348	110
1927	110	4	16	440	114
1928	118	5	25	590	118

$\Sigma Y = 648 \quad \Sigma X = 15 \quad \Sigma X^2 = 55 \quad \Sigma XY = 1,690 \qquad 648$

$M_Y = 108 \quad M_X = 2.5 \quad M_X\Sigma X = 37.5 \quad M_Y\Sigma X = 1,620$

$\Sigma x^2 = 17.5 \quad \Sigma xy = 70$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{70}{17.5} = 4; \quad a = M_Y - bM_X = 108 - 10 = 98$$

$$T = 98 + 4\,X \text{ (origin, 1923)}$$

and subtracted backward as before.   Of course, account is taken of the sign of b in this building-up process.

General solutions.—Occasions often arise where it is not convenient to fit a straight-line trend by the short-cut method utilizing a centered X scale (x) as just described.   This is particularly true when X is not a time series, as in correlation. Hence it may be worth while to consider more general methods of solution.

As has already been suggested, the most direct method of solution is by means of the unmodified normal equations.   Suppose, for example, that for purposes of comparison it had seemed desirable to express the trend in Example 10·3 so that 1923 was the point of origin, that is, so that time was written as $X = 0$, 1, 2, 3, 4, and 5.   In that case the normal equations might be utilized as in part I.   Inspection of the equations will show that numerical values of N, ΣX, ΣY, ΣX^2, and ΣXY are needed. These are readily obtained from the data and substituted. Each equation is then divided by the coefficient of a, and the first equation, thus reduced, is subtracted from the second. Thus it is found that $1\frac{1}{6}b = 4\frac{2}{3}$, and hence $b = 4$.   The value of b may be substituted in the first equation to secure the value of a.   The trend, therefore, is

$$T = a + bX = 98 + 4X \text{ (origin 1923)}$$

The same trend would be obtained by the time-centered method, though the equation then would be

$$T = a + bx = 108 + 4x \text{ (origin 1925.5)}$$

In part II of Example 10·3, a more convenient and useful general method is illustrated.   The procedure may be derived from the normal equations,[1] but it is more easily justified by noting

---

[1] Derivation of b and a from normal equations is readily summarized as follows:

$$Na + b\Sigma X = \Sigma Y \tag{1}$$

$$a\Sigma X + b\Sigma X^2 = \Sigma XY \tag{2}$$

From (1):

$$Na = \Sigma Y - b\Sigma X \tag{3}$$

$$a = \frac{\Sigma Y}{N} - \frac{b\Sigma X}{N} = M_Y - bM_X \tag{4}$$

that b is found as in the time-centering method ($\Sigma xy = \Sigma xY$), and that a is also described as T by that method in the year now taken as the origin. It will be noted that the centering equations described in an earlier chapter are employed, namely,

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

and,

$$\Sigma xy = \Sigma XY - \frac{\Sigma X \Sigma Y}{N}$$

though the correction terms are abbreviated to $M_X \Sigma X$ and $M_Y \Sigma X$ or its equivalent $M_X \Sigma Y$, respectively.

It is important that the general methods just described should be understood, since they will be encountered in expanded forms in later chapters.

READINGS

EXERCISES AND PROBLEMS

A. Exercises

1. By the method of semi-averages, fit straight-line trends to the following annual indexes (consecutive years, as 1931, 1932, etc.). Plot data and trend.

(a)	(b)	(c)	(d)	(e)	(f)	(g)
100	90	172	104	100	95	115
101	88	170	110	92	105	126
108	80	178	109	89	100	119
		172	112	90	108	123
			120	84	112	115
					125	122

From (4):

$$a + bM_X = M_Y \tag{5}$$

Dividing (2) by ΣX:

$$a + b\frac{\Sigma X^2}{\Sigma X} = \frac{\Sigma XY}{\Sigma X} \tag{6}$$

Subtracting (5) from (6):

$$b\frac{\Sigma X^2}{\Sigma X} - bM_X = \frac{\Sigma XY}{\Sigma X} - M_Y \tag{7}$$

Multiplying (7) by ΣX:

$$b\Sigma X^2 - bM_X\Sigma X = \Sigma XY - M_Y\Sigma X \tag{8}$$

$$b(\Sigma X^2 - M_X\Sigma X) = \Sigma XY - M_Y\Sigma X \tag{9}$$

$$b = \frac{\Sigma XY - M_Y\Sigma X}{\Sigma X^2 - M_X\Sigma X} = \frac{\Sigma xy}{\Sigma x^2} \tag{10}$$

2. To the data of Exercise 1, fit least-squares trends and note the resulting change in the value of b. Plot data and trend.

3. By the method of least squares, fit straight-line trends to the following annual index numbers (consecutive years, as 1921, 1922, etc.). Plot data and trend.

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
71	98	92	321	103	65	108	74
95	76	88	288	114	80	106	80
97	86	89	341	112	114	112	78
107	88	90	240	122	96	106	84
85	112	91	200	99	107		88
							82

(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)
90	116	104	120	109	140	119	125
99	110	106	114	111	120	123	119
97	112	101	116	106	170	126	120
106	106	92	110	97	130	124	120
111	102	94	106	99	160	124	122
109	108	85	112	90	110	123	119
					150	129	115

4. Assuming X scale to be 0, 1, 2, etc. (that is, origin in the first year given), recalculate the trend equations of Exercise 3. Use either simultaneous equations (page 236) or the method explained on page 241, part II.

Answers to Exercises

1. (a) $a = 103$ $b = 4$ (e) $a = 91$ $b = -3$
    (b) $a = 86$ $b = -5$ (f) $a = 107.5$ $b = 5$
    (c) $a = 173$ $b = 2$ (g) $a = 120$ $b = 0$
    (d) $a = 111$ $b = 3$

2. (a) $a = 103$ $b = 4$ (e) $a = 91$ $b = -3.4$
    (b) $a = 86$ $b = -5$ (f) $a = 107.5$ $b = 5.114$
    (c) $a = 173$ $b = 0.8$ (g) $a = 120$ $b = .1714$
    (d) $a = 111$ $b = 3.4$

3. (a) $a = 91$ $b = 4$ (i) $a = 102$ $b = 4$
    (b) $a = 92$ $b = 4$ (j) $a = 109$ $b = -2$
    (c) $a = 90$ $b = 0$ (k) $a = 97$ $b = -4$
    (d) $a = 278$ $b = -29$ (l) $a = 113$ $b = -2$
    (e) $a = 110$ $b = 0$ (m) $a = 102$ $b = -4$
    (f) $a = 92.4$ $b = 10$ (n) $a = 140$ $b = 0$
    (g) $a = 108$ $b = 0$ (o) $a = 124$ $b = 1$
    (h) $a = 81$ $b = 2$ (p) $a = 120$ $b = -1$

4. (a) $a = 83$ $b = 4$ (i) $a = 92$ $b = 4$
(b) $a = 84$ $b = 4$ (j) $a = 114$ $b = -2$
(c) $a = 90$ $b = 0$ (k) $a = 107$ $b = -4$
(d) $a = 336$ $b = -29$ (l) $a = 118$ $b = -2$
(e) $a = 110$ $b = 0$ (m) $a = 112$ $b = -4$
(f) $a = 72.4$ $b = 10$ (n) $a = 140$ $b = 0$
(g) $a = 108$ $b = 0$ (o) $a = 121$ $b = 1$
(h) $a = 76$ $b = 2$ (p) $a = 123$ $b = -1$

B. PROBLEMS

5. Following are the numbers of fatalities occasioned by automobiles in the state of Connecticut during day and night hours from 1926 to 1935.

YEAR	DAY	NIGHT	TOTAL
1926	119	195	314
1927	145	196	341
1928	200	220	420
1929	170	270	440
1930	135	260	395
1931	142	300	442
1932	125	255	380
1933	142	305	447
1934	124	325	449
1935	120	320	440

(a) Fit a straight-line trend to these data by the method of semi-averages and the method of least squares.

(b) Chart the data, indicating each of the trends.

(c) In the absence of information relative to factors influencing the trend in 1935 and 1936, would an extrapolation be justified as a means of predicting probable fatalities in 1936?

6. The following data represent the number of banks suspended in the United States for the years indicated. Fit a straight-line trend to these data, and compare this trend with a comparable trend representing business growth.

YEAR	BANKS SUSPENDED
1921	505
1922	367
1923	646
1924	775
1925	618
1926	976
1927	669
1928	499
1929	659

7. The following data represent the annual train-miles of freight traffic on railroads in the United States for a number of years.

YEAR	TRAIN-MILES (in millions)
1911	626.5
1912	612.3
1913	643.8
1914	607.9
1915	552.0
1916	632.3
1917	646.4
1918	628.4
1919	560.5
1920	619.5
1921	519.8
1922	544.5
1923	631.1
1924	590.9
1925	602.9

(*a*) Fit a straight-line trend to these data by the method of semi-averages.

(*b*) Estimate probable normal train-miles for 1933 from the trend thus calculated. (The actual figure for 1933 is 368.7.) Explain the inequality.

8. The following data represent new capital issues in the United States for a period of years.

YEAR	ISSUES (in millions of dollars)
1920	3,634.8
1921	3,576.7
1922	4,304.4
1923	5,593.2
1924	6,220.2
1925	6,334.1
1926	7,791.1
1927	8,114.4
1928	10,182.8
1929	7,023.4
1930	3,115.5
1931	1,192.2
1932	709.5
1933	1,419.5

(*a*) Fit a straight-line trend to these data for the years 1920–1929 by the method of least squares. Chart the data and the trend.

(*b*) May this trend be plausibly extrapolated to represent a normal in 1935?

9. Production (in millions of bushels) of corn, wheat, and oats in the United States, from 1913 to 1938, is summarized in the following table:

Crop year	Corn	Wheat	Oats	Crop year	Corn	Wheat	Oats
1913	2,273	751	1,039	1926	2,547	832	1,153
1914	2,524	897	1,066	1927	2,616	875	1,093
1915	2,829	1,009	1,435	1928	2,666	914	1,313
1916	2,425	635	1,139	1929	2,521	823	1,113
1917	2,908	620	1,443	1930	2,080	886	1,275
1918	2,441	904	1,429	1931	2,576	942	1,124
1919	2,679	952	1,107	1932	2,931	757	1,251
1920	3,071	843	1,444	1933	2,400	552	733
1921	2,928	819	1,045	1934	1,461	526	542
1922	2,707	847	1,148	1935	2,304	626	1,195
1923	2,875	759	1,227	1936	1,507	627	786
1924	2,223	842	1,416	1937	2,651	876	1,162
1925	2,798	669	1,405	1938	2,542	931	1,054

(a) Calculate straight-line trends for each of the three groups as given in the above table by the method of semi-averages and the method of least squares. Plot the data with the latter trend.

(b) Compare the trends thus obtained with similar trends for population in the United States on the basis of census years, 1910–1930, by charting on ratio paper.

10. The following data represent motor vehicle production (in thousands) in the United States.

YEAR	PRODUCTION	YEAR	PRODUCTION
1919	1,934	1925	4,428
1920	2,227	1926	4,506
1921	1,682	1927	3,580
1922	2,646	1928	4,601
1923	4,180	1929	5,622
1924	3,738		

To these data fit a straight-line trend by the method of least squares.

11. Many series of data for which straight-line trends are suitable will be found in the current issues of the *Survey of Current Business*, the *Statistical Abstract of the United States*, the *Monthly Labor Review*, and the *International Labour Review*, as well as in the *Handbook of Labor Statistics*, the annual *Yearbook of Railroad Information*, and the *Yearbook of the New York Stock Exchange*. Obtain suitable data from these or similar sources, and fit straight-line trends by each of the methods described.

CHAPTER XI

COMPLEX TRENDS

Although the straight line is by far the most commonly used type of trend, cases often arise where it is clearly not appropriate. A chart of the data may reveal a gradual change in the direction of trend, or perhaps the trend tends to flatten out as it approaches or reaches a higher or a lower level. In the former case a parabola is suggested, and in the latter an exponential or, more likely, a modified exponential, type of trend. These types of trends will be considered in the present chapter.

THE PARABOLA

The second-degree parabola may be described by the equation,

$$T = a + bX + cX^2$$

But if time centering and a regular sequence of X intervals (a Y for each successive month or year) are assumed, it may be described by the more convenient equation

$$T = a + bx + cx^2$$

where x represents the deviations from the average of the X series. This equation, like that of the straight-line trend, includes an element and a constant for each term. The elements are the successive powers of x, i.e., x^0, x^1, and x^2, which are combined with the constants a, b, and c to define the particular curve that is most nearly representative of the data.[1] The constant a is again a measure of height at the point of

---

[1] The straight-line and parabola belong to a group of trends called the *potential series*, which become increasingly complex by the addition of terms in higher powers of X, as a, bX, cX^2, dX^3, etc. If the maximum power of X is 3, the curve is called a cubic; if 4, a quartic, etc.

origin (where $x = 0$), b is a measure of slope at the origin, and c is an additional increment determining the degree of curvature. The problem involved in fitting a parabola is the discovery of appropriate values for these constants, a, b, and c.

The parabola is most commonly fitted by the method of least squares. In order to facilitate determination of the values of the constants, mathematicians have prepared normal equations similar to those mentioned in connection with the fitting of the straight-line trend. The normal equations required for determining these values in the general case of the parabola are:

$$Na + b\Sigma X + c\Sigma X^2 = \Sigma Y$$

$$a\Sigma X + b\Sigma X^2 + c\Sigma X^3 = \Sigma XY$$

$$a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 = \Sigma X^2 Y$$

But if time centering and a regular x sequence are assumed, then X becomes x, and Σx and Σx^3 will each equal zero. By simple algebraic manipulation, the second equation then provides the value of b as

$$b = \frac{\Sigma x Y}{\Sigma x^2}$$

as in the straight-line trend. Similarly, the first and last equations may readily be reduced to provide formulas for c and a, as follows:

$$c = \frac{N\Sigma x^2 Y - \Sigma x^2 \Sigma Y}{N\Sigma x^4 - \Sigma x^2 \Sigma x^2}$$

$$a = \frac{\Sigma Y - c\Sigma x^2}{N}$$

By substituting the specified summations, which may be readily obtained from the data (as illustrated in Example 11·1), in these formulas, the values of b, c, and a are obtainable. Obviously, it is possible to arrive at the same trend by solving the normal equations in their original form, but the constants a and b will vary with changes in the point of origin in the time scale.

The process of fitting a parabolic trend is illustrated in Example 11·1. The data represent wholesale prices in the United

EXAMPLE 11·1

FITTING A PARABOLA TREND

Data: Index numbers of wholesale prices, United States, 1895–1915 (1913 taken as 100).

Year X	Data Y	Time x	Summed products				a	$+bx$	$+cx^2$	Trend $= T$	Difference Δ_1
			xY	x^2	x^2Y	x^4					
1895	70	-10	-700	100	7,000	10,000				65.212	2.771
1896	67	-9	-603	81	5,427	6,561				67.983	2.667
1897	67	-8	-536	64	4,288	4,096				70.650	2.563
1898	70	-7	-490	49	3,430	2,401				73.213	2.459
1899	75	-6	-450	36	2,700	1,296				75.672	2.355
1900	81	-5	-405	25	2,025	625				78.027	2.251
1901	79	-4	-316	16	1,264	256				80.278	2.147
1902	84	-3	-252	9	756	81				82.425	2.043
1903	86	-2	-172	4	344	16				84.468	1.939
1904	86	-1	-86	1	86	1	88.242	- 1.783	- 0.052 =	86.407	1.835
1905	86	0	0	0	0	0	88.242	0.0	0.0 =	88.242	1.731
1906	89	1	89	1	89	1	88.242	+ 1.783	- 0.052 =	89.973	1.627
1907	94	2	188	4	376	16				91.600	1.523
1908	90	3	270	9	810	81				93.123	1.419
1909	97	4	388	16	1,552	256				94.542	1.315
1910	101	5	505	25	2,525	625				95.857	1.211
1911	93	6	558	36	3,348	1,296				97.068	1.107
1912	99	7	693	49	4,851	2,401				98.175	1.003
1913	100	8	800	64	6,400	4,096				99.178	0.899
1914	98	9	882	81	7,938	6,561				100.077	0.795
1915	101	10	1,010	100	10,100	10,000				100.872	
	1,313		1,373	770	65,309	50,666				1,813.042	

$$b = \frac{\Sigma xY}{\Sigma x^2} = \frac{1,373}{770} = 1.783$$

$$c = \frac{N\Sigma x^2Y - \Sigma x^2 \Sigma Y}{N\Sigma x^4 - \Sigma x^2 \Sigma x^2} = \frac{1,371,489 - 1,396,010}{1,063,986 - 592,900} = \frac{-24,521}{471,086} = - 0.052$$

$$a = \frac{\Sigma Y - c\Sigma x^2}{N} = \frac{1,813 - (-40.0801)}{21} = 88.242$$

States from the low point in 1896 up to the first World War, and it is required to state the general trend or movement of prices during this period. When the data are plotted, a small degree of curvature is apparent, for which reason the parabola is selected as a suitable type of trend (see Fig. 11·1).
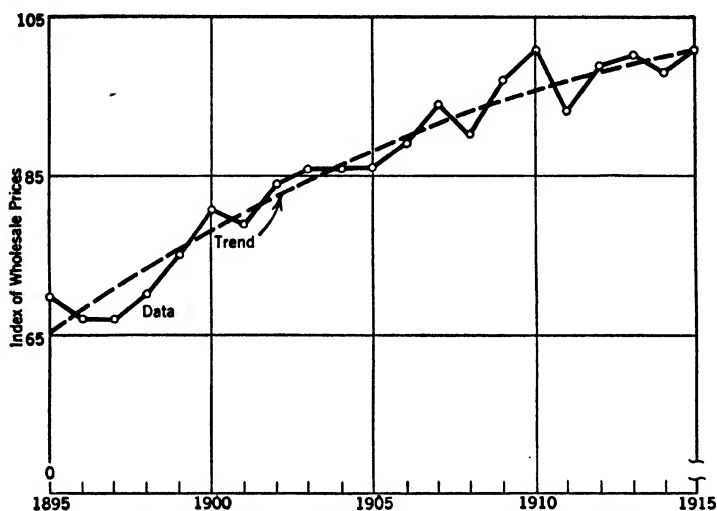

FIG. 11·1.—Least-Squares Parabola Fitted to the Data of Example 11·1.

Building up the parabolic trend.—In computing the trend values for the parabola, the time scale, as has been said, is centered, and the values of ΣxY, Σx^2, $\Sigma x^2 Y$, and Σx^4 are secured, as indicated, for substitution in the equations of the constants. When the substitutions have been made, the trend equation appears as

$$T = 88.242 + 1.783x - 0.052x^2 \text{ (origin at 1905)}$$

This equation may then be solved for all the successive values of x. But in practice, the solution of the trend equation for each time item is generally avoided, and the same result is obtained by a "building-up" process somewhat similar to that described in connection with the straight-line trend. In the case of the parabola, however, the differences between successive trend items, known as *first differences* of the trend (symbol Δ_1), are not equal. *Second differences* (symbol Δ_2), which represent the differences between successive first differences, how-

ever, are the same throughout the series of data (if the time intervals are uniform and the trend values are accurately computed), and they are equal to $2c$. In the illustration, for example, the second differences are 0.104 ($c = 0.052$).

In practice, therefore, it is considerably more convenient to build up the series of first differences by a process of addition, just as a straight-line trend is built up. The process is begun by finding three T values near the point of origin, from which two first differences may be derived (note the italicized items in Example 11·1). The second difference (the difference between the two first differences, in this case $1.835 - 1.731 = 0.104$) is then noted. It is always equal to $2c$. This second difference is then algebraically added forward and subtracted backward from the first differences to secure a series of first differences for the entire period. The trend may then be readily calculated by adding or subtracting these first differences from the trend value at the point of origin, beginning with the central trend items originally determined. Calculation should be carried out to several decimals to avoid cumulative errors, but the final trend items are preferably rounded.[1]

. **Checks.**—The trend thus fitted may be checked in several ways. It will be clear that the sum of the trend items should approximately equal the sum of the data. Also, if the trend is calculated directly from the equation, the second differences, as previously noted, should equal $2c$. With the trend items suitably rounded, however, this check may be only approximate. It is worth noting that these checks cannot prove conclusively the correctness of calculations, although they may catch errors. A further and generally useful check involves the plotting of the data and the trend, and another consists of substituting the constants of the trend equation in the last normal equation, which is

$$a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 = \Sigma x^2 Y$$

---

[1] The calculation of such a trend may be reduced in practice by cumulating the columns xY and x^2Y on the calculating machine without recording each item. The quantities Σx^2 and $(N\Sigma x^4 - \Sigma x^2 \Sigma x^2)$ which are functions of N and x only, and not of the Y's, may be read from a table (see Appendix, page 526), making the column x^4 unnecessary.

If this check is applied to the data of the example, the resulting equation appears as

$$88.242 \times 770 + 1.783 \times 0 - (0.052 \times 50{,}666) = 65{,}312$$

The small discrepancy in this check (65,312 instead of 65, 309) is accounted for by the rounding of the constants to three decimals.

Parabolas fitted by weights.—In shorter problems involving unit x sequence, it is frequently convenient to fit parabolas by especially prepared tables. Simple tables of this type designed to provide values of a, b, and c, and designated values of T, appear in the Appendix, pages 521–524. Required calculations are illustrated in Example 11·2. Each calculation is similar to

EXAMPLE 11·2

PARABOLA FITTED BY WEIGHTS

Data: Assumed index numbers of production (Y) and weights (W) from table, pages 521–522, here designated by the constant or trend item to be calculated. Time unit is here a decade.

Decades	x	Y	W_a	W_aY	W_b	W_bY	W_c	W_cY	W_{T_2}	$W_{T_2}Y$	$W_{T_{-2}}$	$W_{T_{-2}}Y$
1890–1899	−2	82	−3	−246	−2	−164	2	164	3	246	31	2542
1900–1909	−1	94	12	1128	−1	−94	−1	−94	−5	−470	9	846
1910–1919	0	95	17	1615	0	0	−2	−190	−3	−285	−3	−285
1920–1929	1	100	12	1200	1	100	−1	−100	9	900	−5	−500
1930–1939	2	89	−3	−267	2	178	2	178	31	2759	3	267
Divisor and Dividend			35	3430	10	20	14	−42	35	3150	35	2870
			$a = 98$		$b = 2$		$c = -3$		$T_2 = 90$		$T_{-2} = 82$	

Note that the weights for T_{-2} (at $x = -2$) are the same as those given in the table for T at $x = 2$, except that they are written in reverse order.

that involved in finding the mean of grouped data, except that, in the case of b and c, the divisor of ΣwY is not the sum of the weights. In the first table (cf. page 521) the weights to be used with a given N are labeled "a", "b", or "c". In the next table, the weights for finding T at any designated x are similarly listed. They may be recorded successively opposite the Y's, as if they were frequencies. To obtain T for a negative x, weights for positive x, written in reverse order, are used.

An additional use of these weights for T may be noted. If

in a regular series of items one item happens to be lacking, a parabolic interpolation, based on a parabola fitted to the given items, may quickly be made. This result is accomplished by entering the missing Y as 0, subtracting its individual weight from the sum of the weights—the latter being chosen as if to compute the missing item—and finding the weighted average as before. Suppose, for instance, that, in a five-item series, Y at $x = -1$ is unknown. A plausible interpolation might be made by utilizing the reversed $x = 1$ weights ($N = 5$), but subtracting the second weight from the divisor, thus

$$
\begin{array}{lcccccc}
Y & 82 & 0 & 95 & 100 & 89 & \\
W & 9 & (13) & 12 & 6 & -5 & \Sigma W = 35 - 13 = 22 \\
\Sigma WY & 738 + & 0 + & 1{,}140 + & 600 & -445 = 2{,}033 &
\end{array}
$$

Interpolation: $2{,}033 \div 22 = 92.409$.

The Y thus obtained is on the parabolic trend fitted to the *five* items. With adequate data this method provides a convenient and plausible interpolation.

General solution.—Both methods of fitting parabolas described in preceding pages are limited to regular series, such as annual time series. In some circumstances, however, it is necessary to fit parabolas to irregular data, hence a more general method is required. The most direct general method consists of solving algebraically the normal equations (page 249) after calculating and substituting the required summations.[1] But this procedure may be somewhat simplified by centering the data—not only the X's, but the X squares and Y's as well. In so doing it is advisable to designate the X and X^2 series as

---

[1] In outline, T is fitted to Y of Example 11·2 (origin at first Y) by use of the normal equations as follows (see Example 11·3 for summations):

$$
\left.\begin{array}{rcl}
5a + 10b + 30c &=& 460 \\
10a + 30b + 100c &=& 940 \\
30a + 100b + 354c &=& 2798
\end{array}\right\} \quad \text{or} \quad \left\{\begin{array}{rcl}
a + 2b + 6c &=& 92 \\
a + 3b + 10c &=& 94 \\
a + 3.3b + 11.8c &=& 93.26
\end{array}\right.
$$

$$
\left.\begin{array}{rcl}
b + 4c &=& 2 \\
0.3b + 1.8c &=& -0.72
\end{array}\right\} \quad \text{or} \quad \left\{\begin{array}{rcl}
b + 4c &=& 2 \\
b + 5.4c &=& -2.2
\end{array}\right.
$$

$$
\begin{array}{ccc}
1.4c = -4.2 & b - 12 = 2 & a + 28 - 18 = 92 \\
c = -3 & b = 14 & a = 82
\end{array}
$$

X_1 and X_2 to avoid confusion between centered x squares and the squares of centered x. The equations then become

$$b\Sigma x_1^2 + c\Sigma x_1 x_2 = \Sigma x_1 y$$

$$b\Sigma x_1 x_2 + c\Sigma x_2^2 = \Sigma x_2 y$$

The summations may be calculated from the data and centered in the usual way. Thus:

$$\Sigma x_1^2 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{N}$$

$$\Sigma x_2 y = \Sigma X_2 Y - \frac{\Sigma X_2 \Sigma Y}{N}, \quad \text{etc.}$$

For the data of Example 11·3 where the X origin is taken in the first time period, the centered equations become

$$\left. \begin{array}{l} 10b + 40c = 20 \\ 40b + 174c = 38 \end{array} \right\} \quad \text{or} \quad \left\{ \begin{array}{l} b + 4.00c = 2.00 \\ b + 4.35c = 0.95 \end{array} \right.$$

Subtracting the second pair of equations sets $0.35c$ equal to -1.05. Hence $c = -3$. Then $b + 4(-3) = 2$, and $b = 14$, and a may be found by substituting in the first uncentered normal equation

$$Na + b\Sigma X + c\Sigma X^2 = \Sigma Y; \quad \text{or} \quad 5a + 10(14) + 30(-3) = 460$$

so that $a = 82$. It will be seen that this solution does not depend upon the regularity of the data, and it is therefore general. The change in a and b as compared with Example 11·2 is due merely to the change in the time origin.

The student, however, will find it distinctly advantageous to master the form of solution utilized in Example 11·3, since it will be found adaptable to later problems. In this form X and X^2 are set down as two independent series, labeled X_1 and X_2, followed by the dependent series, Y. A check column, Z, is then added, consisting of the row totals. The combined footing of X_1, X_2, and Y should therefore check with the footing of Z. Also, ΣZ^2 and $N\Sigma z^2$ should check as column totals.

In the next four rows—the block P—are set down the summed cross products, that is, the successive sums of each column

multiplied by itself and by each succeeding column. Each sum is designated by the symbols at the left and above, which indicate the columns crossed (including squares). For example, 30 is $\Sigma X_1 X_1 = \Sigma X_1^2$; 100 is $\Sigma X_1 X_2$; 940 is $\Sigma X_1 Y$, etc.

<div align="center">EXAMPLE 11·3</div>

<div align="center">PARABOLA—GENERAL METHOD</div>

Data: See Example 11·2.

	Time	(X^2)	Index	Σ rows	$82 + 14X - 3X^2$
Decades	X_1	X_2	Y	Z	T
1890–1899	0	0	82	82	82
1900–1909	1	1	94	96	93
1910–1919	2	4	95	101	98
1920–1929	3	9	100	112	97
1930–1939	4	16	89	109	90
Sums:	10	30	460	500	460

P	X_1	30	100	940	1,070
	X_2		354	2,798	3,252
	Y			42,506	46,244
	Z				50,566
Np	X_1	50	200	100	350
	X_2	(200)	870	190	1,260
	Y			930	1,220
	Z				2,830

$$50b + 200c = 100; \quad b + 4.00c = 2.00$$
$$200b + 870c = 190; \quad \underline{b + 4.35c = 0.95}$$
$$0.35c = -1.05$$
$$c = -3$$

$$b + 4(-3) = 2; \quad b = 14$$
$$Na = \Sigma Y - b\Sigma X_1 - c\Sigma X_2 = 460 - 140 + 90 = 410$$
$$a = 410 \div 5 = 82$$

In block Np are listed the P's centered and multiplied by N. That is, to avoid fractions, the centering is done by equations like

$$N\Sigma x_1^2 = N\Sigma X_1^2 - \Sigma X_1 \Sigma X_1$$

$$N\Sigma x_2 y = N\Sigma X_2 Y - \Sigma X_2 \Sigma Y$$

The correction term for each sum of cross products is readily located by means of symbols at the left and above. For example, $\Sigma X_2 Y = 2{,}798$ is corrected thus:

$$N\Sigma x_2 y = N\Sigma X_2 Y - \Sigma X_2 \Sigma Y$$

$$= (5 \times 2{,}798) - (30 \times 460) = 190$$

That is, the correction term is the product of the footings of the columns which have been cross multiplied.

The arrangement of the data in blocks P and Np may be summarized in symbols as shown in Table 11·1.

TABLE 11·1

MODEL FORM FOR FITTING PARABOLA

(As applied in Example 11·3)

Headings.........	X_1	X_2	Y	Z
Totals............	ΣX_1	ΣX_2	ΣY	ΣZ
P X_1	ΣX_1^2	$\Sigma X_1 X_2$	$\Sigma X_1 Y$	$\Sigma X_1 Z$
X_2		ΣX_2^2	$\Sigma X_2 Y$	$\Sigma X_2 Z$
Y			ΣY^2	$\Sigma Y Z$
Z				ΣZ^2
Np X_1	$N\Sigma x_1^2$	$N\Sigma x_1 x_2$	$N\Sigma x_1 y$	$N\Sigma x_1 z$
X_2		$N\Sigma x_2^2$	$N\Sigma x_2 y$	$N\Sigma x_2 z$
Y			$N\Sigma y^2$	$N\Sigma y z$
Z				$N\Sigma z^2$

Formulas for centering (Np):

$$N\Sigma x_1^2 = N\Sigma X_1^2 - (\Sigma X_1)^2 \qquad N\Sigma x_2 y = N\Sigma X_2 Y - \Sigma X_2 \Sigma Y$$

$$N\Sigma x_1 x_2 = N\Sigma X_1 X_2 - \Sigma X_1 \Sigma X_2 \qquad N\Sigma x_2 z = N\Sigma X_2 Z - \Sigma X_2 \Sigma Z$$

$$N\Sigma x_1 y = N\Sigma X_1 Y - \Sigma X_1 \Sigma Y \qquad N\Sigma y^2 = N\Sigma Y^2 - \Sigma Y \Sigma Y$$

$$N\Sigma x_1 z = N\Sigma X_1 Z - \Sigma X_1 \Sigma Z \qquad N\Sigma y z = N\Sigma Y Z - \Sigma Y \Sigma Z$$

$$N\Sigma x_2^2 = N\Sigma X_2^2 - \Sigma X_2 \Sigma X_2 \qquad N\Sigma z^2 = N\Sigma Z^2 - \Sigma Z \Sigma Z$$

The final step is the solution of the centered normal equations for b and c, as indicated, together with the solution of the uncentered normal equation for a. The method of solution thus indicated may be expanded for use with any of the potential series, or with any suitable independent series. Later, it will be utilized in multiple and curvilinear correlation.

EXPONENTIAL TRENDS

It is frequently true that the composite direction of change in a series of data throughout a certain period of time may be most effectively represented by an exponential curve, or by some modification of it (see Fig. 10·3b, page 227). Trends of this
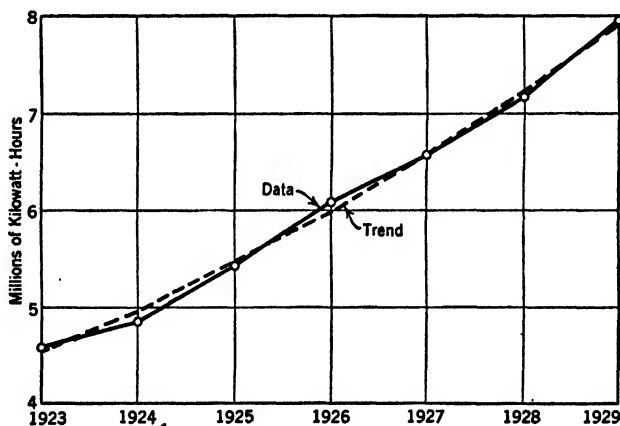

FIG. 11·2.—The Geometric Trend. Trend of electric power production in the United States, 1923–1929. Data: Monthly averages for each year. Source: *Survey of Current Business*, 1938 Supplement, p. 99.

type are characteristic of industries in the early stages of their growth, when they are expanding rapidly. An illustration may be found in the field of electrical power production in the United States in pre-depression years (1919 to 1929), illustrated in Fig. 11·2. The need for this type of trend may be discovered by plotting the original data on ratio or semi-logarithmic paper. If the composite direction of change, whether up or down, appears to approximate a straight line, the *geometric* or *exponential* trend is probably most appropriate for the particular situation.

The exponential trend itself is not often employed, but it is important as an introduction to a group of trends of which the exponential series is an element.   Data of production and of population commonly require a trend belonging to this group.

Fitting an exponential trend.—The method of fitting an exponential trend is not complicated.   A straight-line trend is fitted to the logarithms of the Y items, instead of being fitted to the Y items themselves.   The fitting is carried on exactly as if the logarithms were the original data.[1]   After the trend has been defined in this way, its antilogarithms are used as trend items for the original data.   The various steps in the process are illustrated in Example 11·4.   The five columns between the double rules represent the fitting of the straight-line trend to the logarithms of the Y data, and the last column is merely the antilogarithms of the trend thus defined.

Such a method of trend fitting has a certain analogy to the calculation of the geometric mean.   In both cases, the data are reduced to logarithms, and both geometric mean and trend are distorted when there are any items at or close to zero.[2]   The equation of the trend as fitted to the logarithms is

$$\log T = 3.7762 + 0.0406x$$

but as fitted to the data (see Fig. 11·2) it may be described (making use of antilogarithms) as

$$T = 5{,}973(1.0981)^x$$

The exponential trend by selected points.—The method of fitting the exponential trend just described is somewhat cumbersome, and a short-cut method is usually advisable.   Such a method is illustrated in Example 11·5.   It is well worth mastering because, with certain modifications, it is adaptable to a wide variety of trend-fitting problems.   It is an adaptation of

---

[1] In terms of the original data rather than the logarithms, the equation of the geometric trend is,

$$T = AB^x$$

in which A and B are the antilogarithms of a and b as computed, respectively. A describes the height of the trend at $x = 0$, and B is a constant factor.

[2] For methods of adjusting an approximate trend, see Appendix, page 526.

the method of selected points already described in the early sections of the preceding chapter.

<div align="center">EXAMPLE 11·4</div>

<div align="center">THE EXPONENTIAL TREND</div>

Data: Electric power production (monthly averages), United States, 1923–1929.[1]

Year	Million kilowatt-hours	Log	Time	Summed products		Log trend	Trend of data
		Y	x	x^2	xY	T	
1923	4,571	3.6600	−3	9	−10.9800	3.6542	4,510
1924	4,845	3.6853	−2	4	−7.3706	3.6949	4,953
1925	5,418	3.7338	−1	1	−3.7338	3.7355	5,439
1926	6,088	3.7845	0	0	0.	3.7762	5,973
1927	6,548	3.8161	1	1	3.8161	3.8168	6,558
1928	7,147	3.8541	2	4	7.7082	3.8574	7,201
1929	7,930	3.8993	3	9	11.6979	3.8981	7,909
	42,547	26.4331		28	1.1378	26.4331	42,543

$$a = \frac{\Sigma Y}{N} = \frac{26.4331}{7} = 3.77616$$

$$b = \frac{\Sigma xY}{\Sigma x^2} = \frac{1.1378}{28} = 0.04064$$

Trend of logs: $T = a + bx = 3.7762 + 0.0406x$

Trend of data: $T = $ (antilog 3.7762) (antilog 0.0406)^x

$$= 5,973 \times 1.0981^x$$

The steps in its calculation, as adapted to the problem at hand, are as follows:

(1) Plot the data on ordinary cross-section paper.

(2) Estimate the trend near the beginning and near the end of the series, and mark two points on the trend as thus estimated

[1] *Survey of Current Business*, 1938 Supplement, p. 99.

for specified years.   Read the values of these trend items (P_1 and P_2) on the Y scale.

EXAMPLE 11·5

THE EXPONENTIAL TREND, BY SELECTED POINTS

Data: See Example 11·4.

Year X	Million kilowatt-hours Y	Estimated trend points P	Log P	Completed log series	Antilogs T	Corrections $M_Y - M_T$ $+ (b_Y - b_T)x$ (origin, 1926)		Final T
1923	4,571			3.6593	4,564	−26.86	−36.27	4,501
1924	4,845	5,000	3.6990	3.6990	5,000	−26.86	−24.18	4,949
1925	5,418			3.7387	5,479	−26.86	−12.09	5,440
1926	6,088			3.7784	6,003	−26.86	0	5,976
1927	6,548			3.8182	6,580	−26.86	12.09	6,565
1928	7,147			3.8579	7,209	−26.86	24.18	7,206
1929	7,930	7,900	3.8976	3.8976	7,900	−26.86	36.27	7,909
	$\Sigma = 42,547$		5)0.1986		$\Sigma = 42,735$			$\Sigma = 42,546$
	$M_Y = 6,078.14$		slope 0.03972		$M_T = 6,105$			$M = 6,078$

$$b_Y = \frac{\Sigma_2 - \Sigma_1}{m(N - m)} = \frac{21,625 - 14,834}{12} = 565.92$$

$$b_T = \frac{\Sigma_2 - \Sigma_1}{m(N - m)} = \frac{21,689 - 15,043}{12} = 553.83$$

$$b_Y - b_T = \overline{12.09}$$

(3) Write the logs of P_1 and P_2 in terms of the Y scale and note the number of time units (t) separating them, in this case, 5 years.

(4) Construct a straight-line trend beginning with log P_1 and having a slope of (log P_2 − log P_1) ÷ t, completing the trend for the whole period of time.   This trend will obviously pass through log P_2.   It is comparable to a trend fitted to the logs of the data.

(5) Find the antilogs of the log trend thus found.   This will approximate a trend of the data.   It may be plotted and, if it seems satisfactory, may be taken as final.

(6) If the trend thus found requires adjusting, the following procedure may be adopted.   Add to it a straight-line trend consisting of a constant obtained by taking the mean of the data less the mean of the trend ($M_y - M_t$), and a slope simi-

larly consisting of the slope of the data less the slope of the trend $(b_y - b_t)$. In this correction, X should be centered. The semi-averages method may be employed in determining the slopes. This correction will provide a final trend which approximates the same average and same slope as the data but is otherwise only slightly changed.

The modified exponential trend.—The modified exponential trend is one in which the exponential series is combined with an addend, or some other element, in order to adapt it to the data. Its calculation may be illustrated by the use of production data in a certain industry, by decades from 1850 to 1920, as illustrated in Example 11·6. When these data are plotted they form a curve which has the appearance of an inverted exponential series.

The process of fitting a modified geometric trend is very similar to that described in the last example, except that three points (P_1, P_2, and P_3), widely spaced at equal time intervals, must be estimated and read from the chart. By means of these three points, the base (a) of the exponential series is estimated by the use of the formula

$$a = \frac{P_2^2 - P_1 P_3}{2P_2 - (P_1 + P_3)}$$

As applied to the data of Example 11·6, a is 1050.61.

The central of the three selected points may now be disregarded, and the differences $(P_1 - a)$ and $(P_3 - a)$ may be noted and utilized in further calculations. From this point, the procedure is practically the same as that illustrated in Example 11·5, except that the base, a, which has been subtracted, must later be restored. The logs of $|P_1 - a|$ and $|P_3 - a|$ are noted, and a straight-line log series is calculated to pass through these two figures, the method being the same as that used in simple straight-line trend fitting. The antilogs of this log series are then recorded, and the negative signs of the original $P - a$ are restored. An exponential series is thus obtained which passes through the points $P_1 - a$, $P_2 - a$, and $P_3 - a$. In order to make it approximate the data, it is necessary to add the base, a, which was originally subtracted from the selected points. The result is a trend (column 1,) that presumably fits the data,

EXAMPLE 11·6

THE MODIFIED EXPONENTIAL TREND

Data: Production in a certain industry, by census years.

| Data | | (1) | (2) | (3) | (2ᵣ) | (1ᵣ) | Adjusted T | |
Census year	Y	P	P − a	Log |P − a| completed	Antilog (sign of P − a)	2ᵣ + a T	Correction	Final T
1850	450	$P_1 = 430$	− 620.61	2.79282	− 620.6	430.0	− 6.5	423.5
1860	532			2.68788	− 487.4	563.2	− 5.2	558.0
1870	641			2.58294	− 382.8	667.8	− 3.8	664.0
1880	770	$P_2 = 750$		2.47801	− 300.6	750.0	− 2.5	747.5
1890	820			2.37307	− 236.1	814.5	− 1.2	813.3
1900	848			2.26813	− 185.4	865.2	0.1	865.3
1910	921	$P_3 = 905$	− 145.61	2.16319	− 145.6	905.0	1.4	906.4
1920	935			2.05825	− 114.4	936.2	2.8	939.0

$\Sigma = 5{,}917$
$M_y = 739.62$
$b_y = 70.69$

col. (3): 6) − 0.62963
$\log c = - 0.1049383$
$c = 0.78534$

$\Sigma = 5931.9$
$M_t = 741.49$
$b_t = 69.37$

$- 2472.9$

5917.0

$$a = \frac{P_2^2 - P_1 P_3}{2 P_2 - (P_1 + P_3)} = \frac{750^2 - 430 \times 905}{2 \times 750 - (430 + 905)} = \frac{173{,}350}{165} = 1050.606. \quad (\log 145.61 - \log 620.61) \div 6 = - 0.10494 = \log c.$$

In col. (3), 2.79282 − 0.10494 = 2.68788, etc.

$T = a + b c^x = 1050.61 - 620.61 \times 0.78534^x$ (origin 1850).   Correction $= (M_y - M_t) + (b_y - b_t)x = - 1.87 + 1.32x$ (origin 1885).

assuming that the trend is of a suitable type and that the selected points are carefully chosen.[1]

It will be seen that a technique of the type described consists essentially of: (1) isolating from the selected points a simplified series—in this case an exponential element expressed as a logarithmic straight line, (2) completing this series, and (3) retracing the steps taken in isolating it. This process is suggested, in the example, by the column headings, such as (1), (2), (3), (2_r), and (1_r), where "r" designates the retraced steps. It will be noted that (2_r) is (2) completed, and, similarly, (1_r) is (1) completed. The procedure may appear to be complex, but the steps are readily followed if the logic of the analysis is understood.

If, after plotting the trend with the data, it seems necessary to adjust the trend, the procedure described in connection with Example 11·5 may be followed. That is, the straight line calculated as $(M_y - M_t) + (b_y - b_t)x$, where x represents time centered, may be added to the trend as calculated. This correction series combines the two terms, $a_y - a_t$ and $b_y - b_t$, and its use provides a final trend that closely approximates the same mean and the same slope as the data.

The method of selected points and final trend adjustment thus illustrated is frequently useful because it is adaptable to a wide variety of problems. For example, a parabola may be fitted by a similar method. Three selected points (P_1, P_2, and P_3), separated by a time interval, t, are chosen, and a parabola is fitted to these points. The fitting may.be accomplished by the usual methods or by means of abbreviated formulas for the constants.[2] The parabola thus determined may be adjusted in

---

[1] The equation of the trend is
$$T = a + bc^x$$
where a is the constant subtracted from P, b is $P_1 - a$, c is the antilog of the logarithmic slope, or $(\log \overline{P_3 - a} - \log \overline{P_1 - a}) \div 2t$, where t is the number of time units (decades) from P_1 to P_2 and from P_2 to P_3, and the origin is at P_1.

[2] If the origin of the Y scale is taken at P_2, these equations may be stated in the following form:
$$b = \frac{P_3 - P_1}{2t}$$
$$c = \frac{P_1 + P_3 - 2P_2}{2t^2}$$
$$a = P_2$$

the manner described in preceding paragraphs in order to give it an appropriate mean and slope.

The modified exponential type of trend will be found useful with data that tend to level off, approaching the horizontal, as is typical of growth and production series in their later stages. This tendency is common in the statistics of business.[1]

The Pearl-Reed curve.—If the earlier stages of growth must also be represented, the modified exponential type of trend will be found to be inappropriate in its usual form, and a further variation is necessary to adapt it to the needs of such series. In
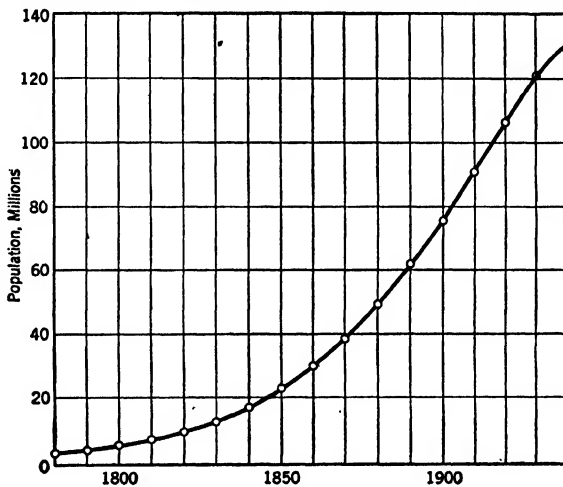
such cases, it is likely that the Pearl-Reed curve or an elementary type of Gompertz curve may be found useful. These curves are so complex that a comprehensive discussion of their characteristics is not possible here, but their general nature may be noted.

The Pearl-Reed curve is illustrated[2] in Fig. 11·3. Several of

[1] For an illustration of a more exact method, see Appendix, page 533.
[2] The Pearl-Reed curve may be stated in equation form as

$$T = \frac{1}{a + bc^x}$$

and the equation of the simple Gompertz curve is

$$T = AB^{c^x}, \quad \text{or} \quad \log T = a + bc^x$$

where $a = \log A$, and $b = \log B$. However, these equations yield the designated curves only when applied to suitable data.

the more complex methods of fitting this type of curve are described in the Appendix (see pages 531 to 533), but it may be fitted with a reasonable degree of accuracy by an adaptation of the technique described in connection with the calculation of the modified exponential trend. The use of this technique in connection with the Pearl-Reed curve is illustrated in Example 11·7. The trend is first fitted to the reciprocals of three selected points, by the means described in Example 11·6. After this fitting has been accomplished, the resulting trend should fit the *reciprocals* of the data. Hence, the reciprocals of this trend may be expected to fit the data in their original form. To avoid inconvenient decimals, reciprocals may be multiplied by a suitable power of 10. If necessary, an adjustment of the level and slope of the curve may be made in the manner already described.

Precautions in trend fitting.—The fitting of trends, perhaps more than any other statistical procedure, requires the use of careful judgment both in selecting the type of trend and in its use. As has already been suggested, it is generally desirable, as a first step, to plot the data and study the chart to see if the points indicate a straight line, a parabola, an exponential, or some other type. Sometimes it may be necessary to experiment with several trends before a satisfactory one can be found.

When a trend has been successfully fitted to a long series of data, it is sometimes extrapolated, that is, calculated for a short distance in advance of the series as a tentative forecast. Such forecasts often prove useful, but they may be very misleading. Unfortunately, no rules can be given to determine whether a forecast thus established has a high or a low probability. In a certain sense a trend is like an established habit; it is likely to continue, but other forces may interrupt or change it. Hence, extrapolation, when used at all, should be for only a comparatively short period of time. An example of fairly satisfactory usage may be found in the estimation of population for a state or nation from one census period until the next census is taken, a process that generally involves no more than the projection of the trend of previous censuses. Such estimates generally prove to be close enough for most practical purposes, although special circumstances may modify their usefulness and accuracy.

EXAMPLE 11·7

THE PEARL-REED (LOGISTIC) CURVE

Data: See Example 11·6.

Census year	Data		(2)	(3)	(4)	(3_r)	(2_r)	(1_r)
	Y	(1) P	Reciprocal (10^4)	Less $a=$ 10.3670	Log	Antilog	Add $a=$ 10.3670	Reciprocal (10^4)
1850	450	430	23.2558	12.8888	*1.11021*	12.8888	23.2558	430.0
1860	532				.89755	7.8986	18.2656	547.5
1870	641				.68488	4.8404	15.2074	657.6
1880	770	750	13.3333	2.9663	.47222	2.9663	13.3333	750.0
1890	820				.25956	1.8179	12.1849	820.7
1900	848				.04689	1.1140	11.4810	871.0
1910	921	905	11.0497	0.6827	-0.16577	0.6827	11.0497	905.0
1920	935				-0.37843	0.4184	10.7854	927.2
	$\Sigma=5917$				6) -1.27598			$\Sigma=5909.0$
	$M_y=739.62$				$\log c = -0.21266$			$M_t=738.62$
	$b_y=70.6875$				$c=0.61283$			$b_t=71.175$

$$a = \frac{(13.3333)^2 - 23.2558 \times 11.0497}{2 \times 13.3333 - (23.2558 + 11.0497)} = \frac{-79.19272}{-7.6389} = 10.3670$$

$$T = 10.3670 + 12.8888 \times (0.6128)^x \quad \text{(origin, 1850)}$$

Trend may be adjusted by adding $(M_y - M_t) + (b_y - b_t)x$ (origin, 1885)

In conclusion, it should be pointed out that the fitting of trends is not simply a mathematical procedure. The element of personal judgment based upon an understanding of the nature of the data and of the problem at hand is always involved. The factors actually represented by the trend are often complex and difficult to analyze. Back of the rising and falling trends of production and prices in our economic system are complex interrelationships involving such varied factors as the advancement of science, changes in monetary units and their values, fluctuations in tariffs and foreign trade, as well as shifts in popular opinions.

On the other hand, there are situations in modern business where the measurement of trends in business data is exceedingly useful if not entirely essential, for some estimate of the future must be made, and no other device ordinarily meets the need as effectively as the fitting and the extrapolation of a trend. Moreover, the trend so defined is continually useful as suggesting what may be regarded as *normal*, thereby permitting valuation of actual data with respect to this criterion. In historical studies, business cycles are measured by this means.

The measurement of trends is applicable not only to series representing business as a whole, such as those typified by the indexes of industrial production and wholesale prices, but also to the less general data of particular localities, industries, and individual firms. Market analysis, for example, may seek to discover trends in such local data as retail sales (in general or with respect to specific goods), bank debits (check transactions), and other business indicators where the composite direction of change appears worthy of consideration. Or the internal statistics of business, such as gross sales, needs in terms of raw materials, labor, and capital, and similar features of the particular firm, may be advantageously examined to discover and measure trends.

READINGS

(See also special and general references, pages 591 and 597.)

Cox, Garfield V., "An Appraisal of American Business Forecasts," *University of Chicago Studies in Business Administration*, 1 (2), 1929.

Davidson, Frederick H., "Interpretations of the Curve of Normal Growth," *Science*, 72 (1861), August 29, 1930, p. 226.

DAVIES, G. R., and CROWDER, W. F., *Methods of Statistical Analysis*, New York, John Wiley & Sons, 1933, Chapter VI.

JOHNSON, NORRIS O., "A Trend Line for Growth Series," *Journal of the American Statistical Association*, 30 (192), December, 1935, pp. 717 ff.

RHODES, E. C., "On the Fitting of Parabolic Curves to Statistical Data," *Journal of the Royal Statistical Society*, 93 (4), 1930, pp. 569–572.

SCHULTZ, HENRY, "The Standard Error of a Forecast from a Curve," *Journal of the American Statistical Association*, 25 (170), June, 1930, pp. 139–185.

STEPHAN, FREDERICK F., "Summation Methods in Fitting Parabolic Curves," *Journal of the American Statistical Association*, 27 (180), December, 1932, pp. 413–423.

WHELDEN, C. H., JR., "Forecast of Automobile Output for 1931–32–33 by a New Method of Analysis," *Annalist*, 37 (952), April 17, 1931, pp. 731–732.

WORKING, HOLBROOK; HOTELLING, HAROLD; and SCHULTZ, HENRY, "The Application of the Theory of Error to the Interpretation of Trends," *Proceedings of the American Statistical Association*, 24 (165A–Supplement) March, 1929, pp. 73–89.

—— "The Determination of Secular Trends," Urbana, Illinois, University of Illinois Bureau of Business Research, June, 1929.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Fit parabola trends to the following data, assumed to represent successive years, by the method of least squares. Results may be checked by Tables A·1 and A·2, pages 521–524.

(a).	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)
10	6	4	6	10	6	26	24	36	44	26	32
16	14	18	30	16	16	16	16	22	20	16	10
34	24	26	40	24	40	34	26	14	10	24	34
24	16	8	16	14	18	40	34	32	34	30	44

(m)	(n)	(o)	(p)	(q)	(r)	(s)	(t)	(u)	(v)	(w)	(x)
9	16	26	21	25	47	49	7	9	9	8	3
8	30	11	22	38	33	32	33	35	18	13	13
15	20	20	15	35	35	33	50	42	19	22	23
20	6	33	10	26	43	42	54	56	22	14	14
13	8	30	17	21	47	49	55	47	23	24	23
							49	41	18	17	17
									17	14	19

2. By the method of least squares fit parabolic trends to the following index numbers representing the items of successive years. Plot data and trend, and as an aid to plotting find the value of x at which the parabola reaches its mode, or mode inverted $\left(\text{maximum or minimum at } x = -\dfrac{b}{2c} \right)$, together with

the height (T) at this point. Check each trend by differencing; the second difference should be a constant.

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)
71	97	94	120	81	91	105	408	136	442	13	9	107	9
95	75	116	96	103	111	130	412	120	498	29	9	108	11
97	85	106	94	103	109	195	440	152	666	23	21	101	19
107	87	104	84	111	115	160	464	176	722	37	24	107	22
85	111	80	106	87	89	165	484	192	582	29	25	105	23
							472	144	652	41	31	116	25
									414	31	21	119	17

3. Calculate the trend items for the data of the preceding exercise using the method of weighted averages of the data explained in the Appendix, page 522.

4. Fit a geometric trend (antilogs of T fitted to logarithms of data) to the following series:

(a) 1, 2, 4, 8, 16, 32, 64.

(b) 19, 25, 30, 32, 43, 50, 57, 75.

The trend will be equal to the data in (a) and approximate it in (b).

5. Plot the following annual index numbers. In each case select three points, P_1, P_2, and P_3, widely separated by t years and estimated to fall upon the trend. Fit a modified geometric trend, and adjust it to the average and slope of the data. Write the equations of the trend.

(a) 203, 212, 210, 250, 300, 340.

(b) 100, 103, 105, 107, 118, 130.

(c) 90, 230, 325, 355, 370, 400.

(d) 70, 82, 90, 98, 97, 100.

(e) 740, 878, 934, 960, 990, 994, 990, 1005, 998.

(f) 1200, 1867, 1880, 1970, 2024, 1967, 1976, 2032, 1990$\frac{5}{6}$.

6. By the method of grouped data, fit modified exponential trends to the data of Exercise 5 (see Appendix, page 533).

<div align="center">ANSWERS TO EXERCISES</div>

1. (a) $a = 26$; $b = 6$; $c = -4$; $T = 8, 22, 28, 26$.

  (b) $a = 20$; $b = 4$; $c = -4$; $T = 5, 17, 21, 17$.

  (c) $a = 24$; $b = 2$; $c = -8$; $T = 3, 21, 23, 9$.

  (d) $a = 38$; $b = 4$; $c = -12$; $T = 5, 33, 37, 17$.

  (e) $a = 21$; $b = 2$; $c = -4$; $T = 9, 19, 21, 15$.

  (f) $a = 30$; $b = 6$; $c = -8$; $T = 3, 25, 31, 21$.

  (g) $a = 24$; $b = 6$; $c = 4$; $T = 24, 22, 28, 42$.

  (h) $a = 20$; $b = 4$; $c = 4$; $T = 23, 19, 23, 35$.

  (i) $a = 16$; $b = -2$; $c = 8$; $T = 37, 19, 17, 31$.

  (j) $a = 12$; $b = -4$; $c = 12$; $T = 45, 17, 13, 33$.

  (k) $a = 19$; $b = 2$; $c = 4$; $T = 25, 19, 21, 31$.

  (l) $a = 20$; $b = 6$; $c = 8$; $T = 29, 19, 25, 47$.

(*m*) $a = 15$; $b = 2$; $c = -1$; $T = 7, 12, 15, 16, 15$.
(*n*) $a = 20$; $b = -4$; $c = -2$; $T = 20, 22, 20, 14, 4$.
(*o*) $a = 20$; $b = 3$; $c = 2$; $T = 22, 19, 20, 25, 34$.
(*p*) $a = 15$; $b = -2$; $c = 1$; $T = 23, 18, 15, 14, 15$.
(*q*) $a = 35$; $b = -2$; $c = -3$; $T = 27, 34, 35, 30, 19$.
(*r*) $a = 35$; $b = 1$; $c = 3$; $T = 45, 37, 35, 39, 49$.
(*s*) $a = 33$; $b = 1$; $c = 4$; $T = 47, 36, 33, 38, 51$.
(*t*) $a = 53$; $b = 8$; $c = -4$; $T = 8, 32, 48, 56, 56, 48$.
(*u*) $a = 50$; $b = 6$; $c = -4$; $T = 10, 32, 46, 52, 50, 40$.
(*v*) $a = 22$; $b = 1$; $c = -1$; $T = 10, 16, 20, 22, 22, 20, 16$.
(*w*) $a = 20$; $b = 1$; $c = -1$; $T = 8, 14, 18, 20, 20, 18, 14$.
(*x*) $a = 20$; $b = 2$; $c = -1$; $T = 5, 12, 17, 20, 21, 20, 17$.

2. (*a*) $a = 103$; $b = 4$; $c = -6$. $Mo = +0.\underline{3}$ $Y_{mo} = 103.\underline{6}$
(*b*) $a = 79$; $b = 4$; $c = 6$. $Mo = -0.\underline{3}$ $Y_{mo} = 78.\underline{3}$
(*c*) $a = 112$; $b = -4$; $c = -6$.
(*d*) $a = 88$; $b = -4$; $c = 6$.
(*e*) $a = 109$; $b = 2$; $c = -6$.
(*f*) $a = 115$; $b = 0$; $c = -6$.
(*g*) $a = 171$; $b = 15$; $c = -10$;
(*h*) $a = 452.5$; $b = 16$; $c = -2$.
(*i*) $a = 165$; $b = 8$; $c = -4$.
(*j*) $a = 680$; $b = 5$; $c = -28$.
(*k*) $a = 33$; $b = 3$; $c = -1$.
(*l*) $a = 24$; $b = 3$; $c = -1$.
(*m*) $a = 105$; $b = 2$; $c = 1$.
(*n*) $a = 22$; $b = 2$; $c = -1$.

3. Same as Exercise 2.

4. (*a*) $a = 0.9031$; $b = 0.3010$; $T = 1, 2, 4$, etc.
(*b*) $a = 1.5778$; $b = 0.0805$; $T = 19.8, 23.8, 28.6$, etc.

5. The trends will approximate those computed in Exercise 6, but may vary owing to choice of selected points.

6. (*a*) $a = 200$; $b = 5$; $c = 2$. (*d*) $a = 100$; $b = -32$; $c = 0.5$.
(*b*) $a = 100$; $b = 1$; $c = 2$. (*e*) $a = 1,000$; $b = -256$; $c = 0.5$.
(*c*) $a = 400$; $b = -320$; $c = 0.5$. (*f*) $a = 2,000$; $b = -729$; $c = \frac{1}{3}$.

B. Problems

7. (*a*) Fit a parabolic trend to the data of Problem 7, Chapter X, page 246.

8. (*a*) Fit a parabolic trend to the data of Problem 8, Chapter X, page 246.

(*b*) May this trend be plausibly extrapolated to represent a normal in 1935?

9. Fit a geometric trend to the following data, which indicate the percentages of population by decades, 1790–1870, living in American cities of 8,000 or more.

1790	3.3 per cent	1840	8.5 per cent
1800	4.0	1850	12.5
1810	4.9	1860	16.1
1820	4.9	1870	20.9
1830	6.7		

10. Following are the approximate numbers of persons (in thousands) moving from city to country and from country to city, 1920–1934.

YEAR	CITIES TO FARMS	FARMS TO CITIES
1920	550	900
1921	750	1,400
1922	1,100	2,200
1923	1,400	2,100
1924	1,550	2,000
1925	1,300	2,000
1926	1,400	2,350
1927	1,700	2,200
1928	1,700	2,150
1929	1,600	2,100
1930	1,750	1,700
1931	1,700	1,450
1932	1,500	1,000
1933	950	1,200
(1934)	(750)	(1,000)

(a) Fit a parabola to each of these series, 1920–1933, inclusive, and plot the data and the trends.

(b) By substitution in the trend equations, estimate the probable numbers moving in each direction in 1934. How close (in per cent) are these predictions to the actual figures.

11. Using the data of Problem 6, Chapter X, page 245, fit the following trends:

(a) A geometric trend by the method of selected points.

(b) A modified geometric trend using the same selected points.

Which type of trend is most appropriate to these figures?

12. To the following index numbers of production in certain industry (base, annual averages 1810–1830) fit a Pearl-Reed curve, and adjust it so that it has the same average and slope as the data. Use the reciprocals (multiplied by 1000), and to these reciprocals fit a modified geometric trend by the method of three selected points.

YEAR	Y	YEAR	Y
1810	29	1870	641
1820	90	1880	770
1830	174	1890	820
1840	302	1900	848
1850	450	1910	921
1860	532	1920	935

13. To the data of Problem 7, Chapter X, page 246, fit a Gompertz curve

employing the same method as in the preceding problem except that the logarithms of the data are used instead of the reciprocals.

14. Many series of data suitable for different types of trends may be found in the *Survey of Current Business*, the *Statistical Abstract of the United States*, and similar sources. Obtain suitable data from these sources, and fit appropriate trends.

CHAPTER XII

SEASONAL VARIATIONS

A survey of time series suggests that they are sometimes characterized not only by long-time secular trends, but also by distinctive patterns of seasonal variation and by longer, more irregular swings generally called *business cycles*. Identification and measurement of these types and patterns of variation represent important features of modern statistical analysis. It is the purpose of this chapter and the next to describe the most common methods by which such variations are discovered and measured.

The seasonal factor.—Investigation of numerous time series discloses a marked tendency of the data to vary according to a fairly regular pattern throughout the year. Certain months or groups of months generally stand well above average levels, and others vary in the opposite direction. For example, agricultural marketings are likely to be large during the fall and small throughout spring months. Similarly, retail trade shows marked advances at Christmas and Easter periods, and numerous special items such as the production of canned goods, the sale of ice and coal, and the packing of fruits have natural seasonal periods. In some cases, seasonal changes in one type of goods or production influence the seasonal fluctuations of other activities from which raw materials are secured or to which finished products are sold. Because of this interrelationship, seasonal variations of one sort or another, and of greater or lesser degree, are found to characterize almost all business series (see, for instance, Fig. 12·1), and their influence extends outside the immediate field of business as well. For example, death rates among industrial workers are affected by the season of the year, usually being high in the winter and low in the summer, and absenteeism and tardiness among such workers show similar seasonal variations.

Indexes of seasonal variation.—Seasonal variations in a particular industry or business are generally expressed in terms of indexes representing the various months or quarters of the year. The average for the whole year is 100, and the individual index numbers indicate the variation above or below that aver-
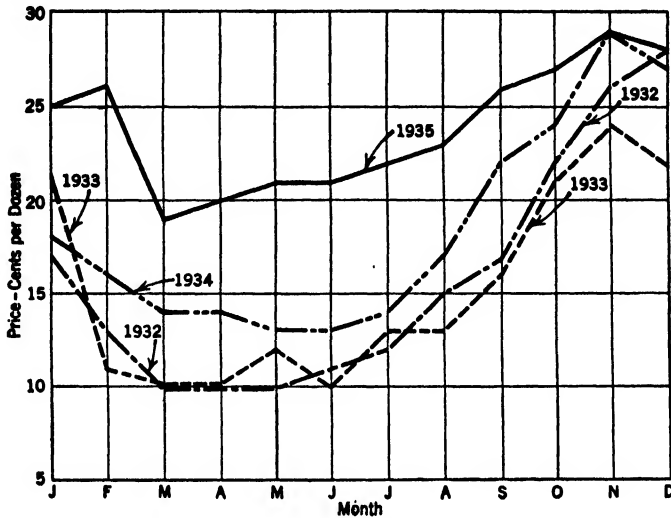

FIG. 12·1.—Seasonal Variability in Farm Prices of Eggs (*Statistical Abstract of the United States*, 1935, p. 606).

age. The characteristics of such a series are obvious from the following monthly indexes representing the seasonal variation in the production of bituminous coal for the years 1928–1931 (Federal Reserve Board indexes):

January	111	July	91
February	106	August	98
March	100	September	106
April	84	October	110
May	87	November	112
June	89	December	106

The index clearly indicates that production of bituminous coal in the United States was generally at low ebb during the summer and became increasingly active as fall and winter progressed.

The variety and usefulness of such seasonal indexes is suggested by the summary in Table 12·1, which shows the indexes

TABLE 12·1
SEASONAL INDEXES *

Series	Indexes					
	Jan.	Feb.	Mar.	April	May	June
Total car loadings	90.2	99.3	94.8	94.1	99.6	100.7
Miscellaneous car loadings	81.2	91.8	93.8	100.5	102.8	104.2
Electric power production	102.4	101.5	99.1	98.1	98.3	98.1
Steel ingot production	103.0	112.3	115.9	112.0	107.6	99.5
Pig iron production	95.2	103.5	107.6	110.2	110.7	104.9
Cotton consumption	104.6	112.7	103.3	103.9	106.0	94.8
Wool consumption	100.7	111.7	94.7	93.0	96.0	96.2
Silk consumption	114.0	116.7	105.2	94.6	93.6	86.0
Rayon production	104.3	108.5	100.1	93.4	92.5	86.2
Boot and shoe production	85.7	111.8	104.2	104.3	98.1	97.4
Automobile production	101.9	103.7	111.0	124.2	123.2	117.5
Lumber production	79.8	87.6	97.9	107.4	110.1	107.8
Cement production	60.4	64.3	69.9	94.3	120.7	125.7
Zinc production	102.3	105.1	105.8	100.9	98.1	96.9
Lead production	103.5	102.8	99.1	99.5	101.4	95.5

Series	Indexes					
	July	Aug.	Sept.	Oct.	Nov.	Dec.
Total car loadings	100.7	104.4	111.8	114.6	99.9	89.9
Miscellaneous car loadings	103.8	106.1	115.2	119.6	99.6	81.4
Electric power production	99.1	98.2	100.0	100.4	102.3	102.5
Steel ingot production	94.6	93.0	92.1	93.7	89.1	87.2
Pig iron production	97.4	97.2	94.1	94.0	93.1	92.1
Cotton consumption	86.3	87.2	96.6	105.3	106.4	92.9
Wool consumption	97.7	103.2	104.6	110.6	102.1	89.5
Silk consumption	91.3	97.1	104.9	104.6	101.6	90.4
Rayon production	94.3	111.0	118.3	103.3	96.5	91.6
Boot and shoe production	99.0	113.8	115.3	111.4	85.3	73.7
Automobile production	108.8	68.6	25.2	63.3	126.0	127.0
Lumber production	107.4	106.2	106.0	103.9	98.0	88.4
Cement production	124.4	121.5	127.1	114.3	100.1	77.3
Zinc production	92.9	95.7	98.3	100.2	102.2	101.6
Lead production	91.4	94.8	96.9	108.5	106.1	100.5

* From the *Annalist*, 47 (1223), June 26, 1936, p. 940, by permission.

formerly used by the *Annalist* in compiling its index of business activity. It will be noted that there is wide variability among the several industries, and the student will find it both interesting and illuminating to chart several of the series in order to facilitate their comparison.

Editing the data.—It is frequently necessary, in preparing such an index of seasonal variation, to edit and adjust the data upon which such calculations are based. Monthly production or sales data, for instance, are often revised to take account of the number of days or the number of working days in each month, thus improving the accuracy of month-to-month comparisons. Similarly, data that express values or prices in dollars may be deflated by suitable indexes of living costs, wholesale prices, or other measures of changing price levels, in order to stress variations in quantity to the exclusion of price changes. In measuring the business cycle, as will be noted in the next chapter, both factors are usually involved, so that value data are often used without deflation. Other types of adjustment may be suggested by the special circumstances of the problem at hand.

Measuring seasonal change.—Methods of calculating seasonal relatives may be applied in the same manner to the crude data of prices and production or to the index numbers representing such data. In some cases, the methods are applied directly to indexes representing aggregates of the data. In this chapter, for instance, the data chosen to illustrate seasonal and cyclical analysis are composite index numbers, but the methods described are equally applicable to data expressed in monetary or physical units.[1]

Indexes of seasonal variation may be calculated by several methods, but the most common one is based upon an annual moving average. If quarterly data are used, the moving average covers 5 items, giving half weight to the extremes. Similarly, if monthly data are used, each moving average may

[1] In many cases, the propriety of applying seasonal and cyclical analysis to composite data in which seasonal fluctuations of individual items vary widely may well be questioned. In the case of retail sales, however, the composite is generally treated as a unit, and the purpose here is only to explain the methodology commonly utilized.

SEASONAL INDEXES, MOVING AVERAGE METHOD

Data: Department-stores sales, United States, 1925–1929 (1932 Annual Supplement, *Survey of Current Business*, pages 48–49).

I. Index numbers (Y), monthly averages 1923–1925 taken as 100.

Time	1925	1926	1927	1928	1929
January	84	90	91	91	90
February	85	87	89	88	91
March	94	97	95	97	107
April	105	102	109	105	103
May	103	109	105	107	109
June	98	100	101	102	108
July	75°	77	76	80	79
August	76·	82	85	81	84
September	97°	104	103	113	117
October	122	120	117	118	122
November	122	124	126	125	125
December	176	184	182	192	191
	1,237	1,276	1,279	1,299	1,326

II. Moving average (MA), 12 months, centered.

Time	1925	1926	1927	1928	1929
January	.....	104.5	106.7	106.8	109.7
February	.....	104.8	106.8	106.8	109.8
March	.....	105.4	106.9	107.0	110.1
April	.....	105.6	106.7	107.5	110.4
May	.....	105.6	106.7	107.5	110.6
June	.....	106.0	106.7	107.8	110.5
July	103.3	106.4	106.6	108.2	.....
August	103.7	106.5	106.5	108.3	.....
September	103.9	106.5	106.6	108.8	.....
October	103.9	106.7	106.5	109.2	.....
November	104.0	106.8	106.4	109.2	.....
December	104.3	106.7	106.5	109.5	.....

III. Seasonal relatives or percentages $\left(SR = \dfrac{Y}{MA} \right)$, and seasonal indexes.

Time	1925	1926	1927	1928	1929	Crude index (Md)	Index (S)
January	.....	86.1	85.3	85.2	82.0	85.25	85.6
February	.....	83.0	83.3	82.4	82.9	82.95	83.3
March	.....	92.0	88.9	90.7	97.2	91.35	91.7
April	.....	96.6	102.2	97.7	93.3	97.15	97.5
May	.....	103.2	98.4	99.5	98.6	99.05	99.4
June	.....	94.3	94.7	94.6	97.7	94.65	95.0
July	72.6	72.4	71.3	73.9	....	72.50	72.8
August	73.3	77.0	79.8	74.8	....	75.90	76.2
September	93.4	97.7	96.6	103.9	....	97.15	97.5
October	117.4	112.5	109.9	108.1	....	111.20	111.6
November	117.3	116.1	118.4	114.5	....	116.70	117.1
December	168.7	172.4	170.9	175.3	....	171.65	172.3
						12)1,195.50	12)1,200.0
						$M = 99.625$	100.0

cover 13 items, giving half weight to the first and last. This approach may be approximated, however, by taking a 12 months' moving average centered in the seventh month of each group. The original data are compared with the moving average by expressing them as percentages of the latter, and these percentages or *seasonal relatives* extending over a period of years are averaged for like quarters or months. The resulting indexes,
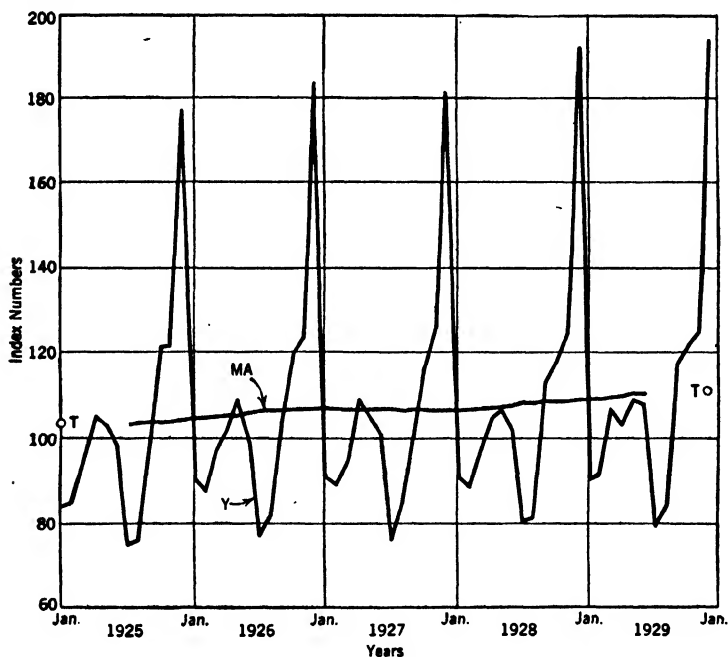

Fig. 12·2.—Index Numbers of Department-Store Sales, 1925–1929 (monthly average 1923–1925 = 100), and Moving Averages.

reduced if necessary so that their own average is the base (100), represent the index of seasonal variations. This method of calculation is illustrated in Example 12·1. The data employed in that example together with the moving average are charted in Fig. 12·2.

Several details of the procedure in Example 12·1 deserve brief explanation. In the first place, it should be observed that each of the moving averages covers exactly one year, in spite of

the fact that each includes 13 items. The calculation in effect takes half-month units, inclusive, from the middle of the first month to the middle of the like month a year later, on the assumption that the data are uniform within the month. As a result, the average is centered in the seventh month of the 13 months nominally included. Thus the first moving average in Example 12·1 covers the year from January 15, 1925, to January 15, 1926, and it is obtained as follows:

$$(84 + 2 \times 85 + 2 \times 94 + 2 \times 105 + 2 \times 103 + 2 \times 98$$
$$+ 2 \times 75 + 2 \times 76 + 2 \times 97 + 2 \times 122 + 2 \times 122$$
$$+ 2 \times 176 + 90) \div 24 = 2{,}480 \div 24 = 103.3$$

It centers at the middle of July, 1925. The second moving average may be similarly obtained by beginning with February, 1925, and continuing through February, 1926, thus,

$$(85 + 2 \times 94 + 2 \times 105, \text{ etc., to } + 87) \div 24$$

In practice, it is most convenient to obtain these indexes by totaling the first group on an adding machine and taking the subtotal. From this subtotal, subtract the first and second items, 84 and 85, and add figures for the corresponding months in the next year, namely 90 and 87, thus:

$$2{,}480 - 84 - 85 + 90 + 87 = 2{,}488$$

Again a subtotal may be taken, and 2 items—February and March—subtracted and added as before, moving down 1 place in the data. This process may be continued to the end of the series; an independent calculation will check the result. After all the subtotals have been obtained, they may be divided by 24, or, in the case of quarterly data, where 5 items are similarly grouped, by 8. The successive quotients constitute the moving averages.

Although the method of calculating the moving average just described is theoretically appropriate, as suggested above, it may usually be materially simplified for monthly data with only a negligible loss of accuracy. This result is accomplished by centering a *12-month* moving average on the seventh month. Thus, for instance, the average of the first 12 months (January–

December), is centered at the seventh month (July). The process is repeated, beginning each average 1 month later and moving the center 1 month later also. In practice the second moving total may be obtained from the first by subtracting the first item and adding the thirteenth. Similarly, succeeding totals may be obtained. Thus for the data at hand, the July, 1925, moving average is

$$[(84 + 85 + 94 + 105 + 103 + 98 + 75 + 76 + 97 \\ + 122 + 122 + 176) \div 12] = [1{,}237 \div 12] = 103.1$$

and the next (August) moving average is

$$(1{,}237 - 84 + 90) \div 12 = 1{,}243 \div 12 = 103.6$$

Similarly, $1{,}243 - 85 + 87$ is the next total. If an adding machine is used, the successive moving totals may be obtained as subtotals, and after checking, each may be divided by 12 (or multiplied on the calculator by the reciprocal of 12, which is 0.0833). This procedure is more economical, and the small errors resulting are largely removed in a final adjustment of the index, to be explained shortly.

Seasonal relatives.—The next step in the calculation of the seasonal index is the comparison of the data for each month with the ordinate of the moving average for the same month to secure monthly relatives. The moving average itself has largely escaped seasonal influence, since each item is the average of an entire year and it may, under certain circumstances to be noted later, be regarded as representing the data corrected for seasonal influences. However, if seasonal variability is at all regular or if it changes with some degree of regularity, it is more satisfactory to express this seasonal variation in the form of seasonal relatives. Furthermore, the annual moving average necessarily lacks 6 months at the end of the series and so cannot be applied to the last or current items. It is for these reasons that the data are generally revalued as ratios to their corresponding moving averages, in order to discover the extent to which seasonal and random influences have caused them to vary.

After the seasonal relatives ($Y \div MA$) have been thus calculated, the averages for each month (the Januaries, Febru-

aries, etc.) are required. In reasonably regular data, these averages will tend to eliminate the specific fluctuations and irregularities of individual months and will thus measure the normal or expected seasonal influence. The kind of average to be employed, however, calls for careful consideration and judgment. Obviously, the arithmetic mean may not be suitable, because it is readily disturbed by occasional erratic items which should not be given full weight. Hence, as a general rule, the median, or some modification of it, is employed. It is expedient not to follow any fixed rule, however. When the data are adequate the median may be modified to comprise an average of several of the central items (usually about a third), thus giving it greater stability. For example, suppose that in a certain problem the following seasonal relatives represented the month of January:

Years	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929
January, SR:	85	82	86	91	88	93	93	92	89	96

These seasonal relatives, arrayed in order of size from the smallest to the largest, are

$$82, 85, 86, 88, 89, 91, 92, 93, 93, 96$$

Since in this case the 4 central items appear to be fairly constant, it might be justifiable to average them, eliminating 3 items at each extreme of the array. Obviously, it is necessary that the same number should be eliminated at each extreme. Hence the typical seasonal is [1]

$$(88 + 89 + 91 + 92) \div 4 = 90.0$$

It will be noted that this average may be easily obtained without arraying the items merely by crossing out the 3 largest and the 3 smallest relatives, and averaging the others.

As a final step, the crude index must be adjusted, if necessary, to make the average (the base of the index) exactly 100.

---

[1] A weighted average of the arrayed relatives stressing the central items is sometimes preferred. In the case cited the weights might be: 0, 0, 1, 2, 3, 3, 2, 1, 0, 0, respectively. These weights result in an average of 89.9, almost identical with that obtained above.

This is necessary so that when it is used later as a divisor of the data, Y, it will not materially change its aggregate level. This adjustment follows the usual method of changing the base of an index series, that is, each item of the index is divided by the average of all the items, as indicated in Example 12·1. The resulting index may be rounded to even percentage figures, or perhaps to one decimal,[1] as here illustrated.

While there are several refinements of the foregoing process, as will be noted later, the index thus obtained is commonly used as a means of adjusting original data for seasonal fluctuations, as is explained in later paragraphs.

The scatter of seasonal relatives.—When sufficient data are at hand, it is desirable to tabulate the seasonal relatives in such a way as to make clear their distribution. This tabulation may be effected by arranging suitable class intervals and limits and entering items in a frequency distribution. Or the same objective may frequently be more conveniently attained by graphic methods, as is illustrated in Fig. 12·3. In this chart, individual seasonal relatives are connected by lines in the space allotted to each month, their sequence indicating consecutive years. The scatter of the dots within each month as compared with the scatter throughout the year suggests that seasonal influences are decisive and reasonably uniform within the limits of the data included in the study. Methods of estimating the reliability of the seasonal scatter will be considered later.

Adjusting for seasonal variations.—Data are said to be adjusted for seasonal variations when they have been divided by an appropriate seasonal index. Each January item of the data is divided by the January seasonal index, each February item by the February seasonal index, etc. As a consequence

[1] In rounding the relatives it may be necessary to make the error thus occasioned more than 0.05. For example, suppose that, in quarterly data, the seasonal indexes, after division by the average, are 92.64, 92.59, 103.39, and 111.33. If they are rounded to one decimal by the usual method, the total is 0.1 low. Hence, 0.04 may be taken instead of 0.05 as the lower limit for adding 0.1. The series then becomes 92.7, 92.6, 103.4, and 111.3 ($\Sigma = 400$). If this approximation is not satisfactory, the earlier computations may be carried out farther and the index rounded to two decimals. In any case, it is desirable that the final index shall average exactly 100. It should also be noted that, even though this average is taken as the base, the data adjusted for seasonal (Y/S) will not necessarily remain unchanged in the aggregate.

the low months are increased and the high months are decreased. If the seasonal is perfectly regular for each year, this procedure eliminates practically all seasonal variation, and the data appear as a relatively smooth line like an extended moving average.

Many series of data commonly used as indicators of business are published in both the unadjusted and adjusted forms. For example, in the *Survey of Current Business*, the Federal Reserve Board index of industrial production is listed under these two
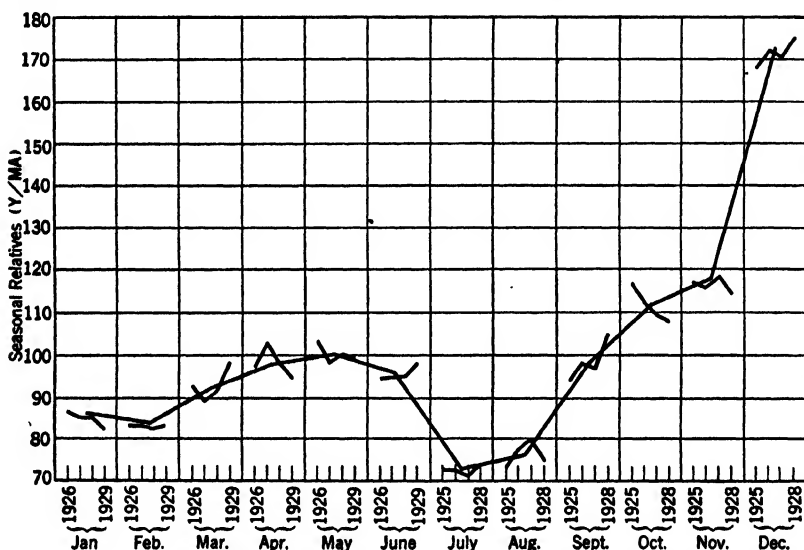

Fig. 12·3.—Seasonal Relatives of Department-Store Sales, July, 1925 to June, 1929. Short lines show scatter of seasonal relatives. Continuous line connects monthly indexes. For data, see Example 12·1.

headings. In the latter case the original series of data have been divided by their seasonal indexes. It will be seen that the advantage of this adjustment is to make unlike months comparable. Thus, in January, 1935, the unadjusted index was 88 and in February it was 91. This was a rise, but not so much as would be seasonally expected, for seasonally adjusted figures for the same months are 91 for January and 89 for February. The adjustment for seasonal is illustrated in Example 12·2, the data of which are the same as in Example 12·1. Although this adjustment is clearer as applied to index numbers, in which case

<div align="center">

EXAMPLE 12·2

THE ADJUSTMENT OF DATA FOR SEASONAL VARIATIONS

</div>

Data: Department-store sales, United States, 1925–1929 (see Example 12·1).

I. Index numbers (Y), monthly averages, 1923–25 = 100

Time	1925	1926	1927	1928	1929
January	84	90	91	91	90
February	85	87	89	88	91
March	94	97	95	97	107
April	105	102	109	105	103
May	103	109	105	107	109
June	98	100	101	102	108
July	75	77	76	80	79
August	76	82	85	81	84
September	97	104	103	113	117
October	122	120	117	118	122
November	122	124	126	125	125
December	176	184	182	192	191
	1,237	1,276	1,279	1,299	1,326

II. Seasonal index (S) (see Example 12·1)

January	85.6	May	99.4	September	97.5
February	83.3	June	95.0	October	111.6
March	91.7	July	72.8	November	117.1
April	97.5	August	76.2	December	172.3

III. Index of department-store sales, adjusted for seasonal variations $(Y \div S)$

Time	1925	1926	1927	1928	1929
January	98.1	105.1	106.3	106.3	105.1
February	102.0	104.4	106.8	105.6	109.2
March	102.5	105.8	103.6	105.8	116.7
April	107.7	104.6	111.8	107.7	105.6
May	103.6	109.7	105.6	107.6	109.7
June	103.2	105.3	106.3	107.4	113.7
July	103.0	105.8	104.4	109.9	108.5
August	99.7	107.6	111.5	106.3	110.2
September	99.5	106.7	105.6	115.9	120.0
October	109.3	107.5	104.8	105.7	109.3
November	104.2	105.9	107.6	106.7	106.7
December	102.1	106.8	105.6	111.4	110.9

the adjusted data may be regarded as percentages of like months in a typical base year, it will, in any case, serve the purpose of comparability as just explained. A comparison of sales and seasonally adjusted sales in recent years is made in Fig. 12·4.
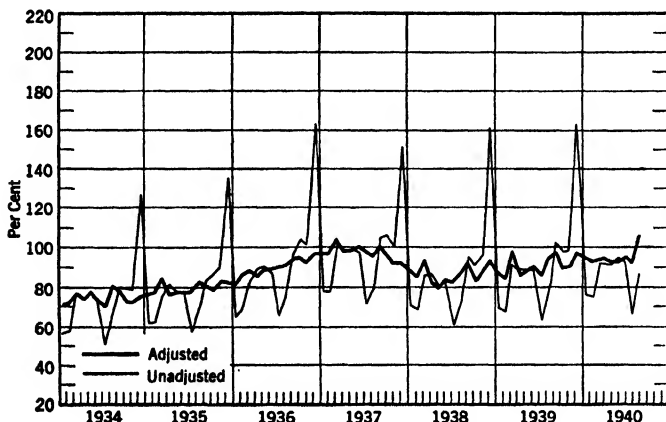

Fig. 12·4.—Indexes of Department-Store Sales in the Seventh Federal Reserve District, with and without Adjustment for Seasonal Variation, by months, 1934 through March, 1940. (Average 1923–24 = 100.) Source: Chicago Federal Reserve Bank.

Final adjustment for seasonal.—When the data have been thus adjusted for seasonal, an additional final adjustment may be desirable. If a very high degree of accuracy is required, it might even be worth while to repeat the whole process of measuring the seasonal, in which case, if the work has been entirely satisfactory, the resulting seasonal index would theoretically be 100. Practically, such a procedure would be uneconomical, but an approach may be made to it by inspecting a chart of the adjusted data for seasonal effect. If, for example, it is found that all the Januaries are slightly higher than the contiguous adjusted figures, then it may be inferred that the seasonal index for January was a little too low, and an adjustment may be estimated accordingly. By this means the effects of inaccurate choice of medians in the computing of the relatives, and in the process of adjusting them, may be reduced.

REFINEMENTS OF SEASONAL MEASUREMENTS

Thus far, in this chapter, the measurement of seasonal variability has been described in terms of the construction of indexes to be applied uniformly to the data upon which they are based. Such a measurement assumes that seasonal influences are practically constant from year to year and that divergences from this regularity are due to erratic influences in particular months. This assumption, however, may be advantageously modified to take into account various factors that cause the seasonal variability to change from month to month or from year to year.

Number of working days in the month.—It has already been suggested that the original data might be rendered more comparable from month to month if account were taken of the number of working days in each month, that is if the data were divided by a figure representing the number of calendar days less the number of holidays. Such an adjustment would put the data on an average workday basis, and the variability in the length of the month would thus be discounted from the start.

Of course, this reduction of the data to a workday basis is not universally applicable. Price indexes, for example, are usually computed as of the first or fifteenth of the month, and variation in the length of the month is of little consequence. The sale of groceries is only slightly affected by holidays, although it is affected by the varying number of calendar days. In this case, however, an adjustment for the number of calendar days would not be necessary, inasmuch as this variation would be reflected in the seasonal index itself and would, therefore, be removed with the adjustment of the data for seasonal variations. The sole exception to this principle would arise out of the additional day in leap years, which would introduce a slight variability not only in the month of February but also in the total for the year.

Whether department-store sales should be adjusted for the number of working days in the month is uncertain. If great exactness is required, it is necessary to deal separately with the various departments. But, from the practical point of view, it is sufficient to try out such an adjustment and to make such

changes as appear justified by the results thus obtained. The Federal Reserve Board, after careful analysis of the figures it compiles, has concluded that such adjustment is advisable.[1]

In the study here presented, however, this adjustment has been omitted, the assumption being that the varying number of calendar days alone is important and that this variability is reflected in the seasonal index. The influence exerted by the varying number of holidays remains in the data and may be adduced to explain the variability of the adjusted data for specific months. Of course, many minor factors that cannot practically be measured exert an influence, such as those affecting the income of customers. Only when such factors become pronounced can they be dealt with, as in the case of the changing date of Easter.

The variability of Easter.—Retail trade is modified in the spring by the changing date of Easter, which varies from March 22 to April 25. The relative volume of trade in April as compared with March will be found to increase as Easter is delayed.

The relation of Easter to retail trade may be readily measured by comparing for each year the variability of trade in

---

[1] When cyclical change is rapid, the moving average may not furnish an adequate base from which to measure seasonal variations. It tends "to cut corners," that is, it does not swing out as far as the real center of the data. Except in extremely erratic cases, however, it may be adjusted by the simple process of computing from it (MA_1) a second moving average (MA_2) covering the same number of items and deriving a final figure for each month as $2MA_1 - MA_2$ on the assumption that the second moving average "cuts corners" about as much as the first. For example, suppose that in an historical study of interest rates prior to the Federal Reserve system it appeared desirable to adjust the moving average of quarterly rates during a minor depression period. The following moving average items would be adjusted as shown (quarters indicated by subscripts):

Year and quarter:	1910_3	$'10_4$	$'11_1$	$'11_2$	$'11_3$	$'11_4$	$'12_1$	$'12_2$	$'12_3$	$'12_4$ ...
MA_1:	5.02	4.80	4.48	4.18	4.02	4.08	4.26	4.59	4.98	5.34
MA_2:			4.50	4.28	4.16	4.19	4.36	4.64		
$2MA_1 - MA_2$:			4.46	4.08	3.88	3.97	4.16	4.54		

The adjustment covers the depression period. This adjustment tends to increase the amplitude of the moving-average cycle.

More commonly, however, such adjustments are made informally by charting the data and the usual moving average, and then, at points where there are reversals of trend, drawing freehand the corrected moving average. The values of this corrected moving average are then read from the chart.

the two months in question as against the date of Easter. This comparison may be made by subtracting from each April seasonal relative the prior March seasonal relative and plotting the results with the date of Easter as the X scale, as in Fig. 12·5. For this purpose, March and April relatives for 1925 have been included. From the data thus extended a straight-line trend has been calculated. It will be seen from the chart that such a trend expresses the relationship fairly well. When the study
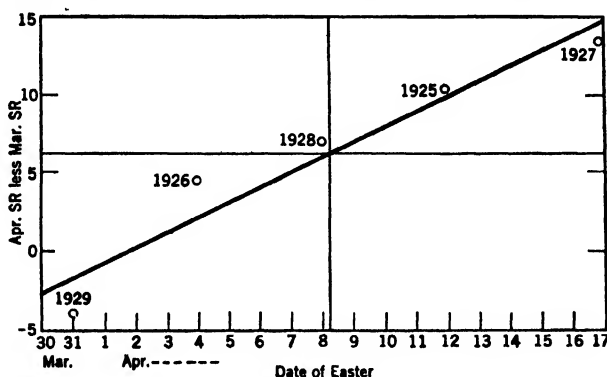

FIG. 12·5.—Trend or Regression of April Seasonal Relatives Less March Seasonal Relatives on Changing Dates of Easter. Department-store sales, 1925–1929.

is extended to earlier years, much the same regression is discovered. Hence, the process may be regarded as reliable enough for practical purposes.[1]

In order to make use of the measurement of the Easter factor, it is necessary to correct the seasonal index for each year. Thus in 1925, $bx/2 = 1.8$ (see Example 12·3) is added to the April item and subtracted from the March item, thus making the index conform theoretically with the later date of Easter in that year. The $\frac{1}{2}bx$ thus applied twice makes, of course, a full bx correction in the April seasonal relative less the March seasonal relative. Similarly, for each year, the $bx/2$ of that year is added to the April item and subtracted from the March item, allowing for the algebraic sign. In this way the seasonal

[1] On the assumption that each Y of Example 12·3 represents an independent case, the reliability may be estimated in terms of a correlation coefficient as explained on page 331.

index is made to vary from year to year with respect to the months of April and March. When the seasonal index thus corrected is used in adjusting the data, the rather extreme vari-

<div align="center">EXAMPLE 12·3</div>

VARIABILITY IN SEASONAL RELATIVES DUE TO THE CHANGING DATE OF EASTER

Year	Seasonal relatives		Increase	Date of Easter	Time		Adjustment	Trend T
	March	April	Y		X	x	$\frac{b}{2}x$	T
1925	93.6*	103.9*	10.3	Apr. 12	12	3.8	1.8	9.86
1926	92.0	96.6	4.6	Apr. 4	4	−4.2	−2.0	2.28
1927	88.9	102.2	13.3	Apr. 17	17	8.8	4.2	14.59
1928	90.7	97.7	7.0	Apr. 8	8	−0.2	−0.1	6.07
1929	97.2	93.3	−3.9	Mar. 31	0	−8.2	−3.9	−1.51
			5)31.3		5)41	0	0	31.29
			$a=6.26$		8.2			
	* New data							

$$\text{Slope: } b = \frac{\Sigma xY}{\Sigma x^2} = \frac{167.44}{176.80} = 0.9471$$

$$\frac{b}{2} = 0.4735$$

ability of March and April will largely disappear, as is evident below. The corrected seasonal index items (S_c) for March and April appear as follows:

Month	1925	1926	1927	1928	1929
March	89.9	93.7	87.5	91.8	95.6
April	99.3	95.5	101.7	97.4	93.6

and, when these items of the seasonal index are used to adjust the original data for March and April, the following corrected indexes $(Y \div S_c)$ are obtained:

Month	1925	1926	1927	1928	1929
March	104.6	103.5	108.6	105.7	111.9
April	105.7	106.8	107.2	107.8	110.0

The adjusted data for March and April are now much closer than they were in the original calculation (see Example 12·2). The sum of the differences, taken as positive, is now 9.8 points; previously, it was 27.6 points. It seems, therefore, that the major portion of the irregularities as between March and April in the earlier method of adjustment was due to the changing date of Easter, and these irregularities have now been approximately allowed for and removed.[1]

Changing seasonals.—A problem that often arises in connection with the measurement of seasonal variation is that of a gradual change from year to year. There are many types of causes that may occasion such a shift in seasonality. The development of all-weather roads, for instance, has had the effect of spreading the sales of automobiles throughout the year and permitting the regularization of production to avoid the concentration of both sales and production that formerly characterized the spring and early summer. This shift, as it has affected production, is graphically shown in Fig. 12·6, which compares monthly production of cars in 1920 and 1922 with that of 1935. But this shift
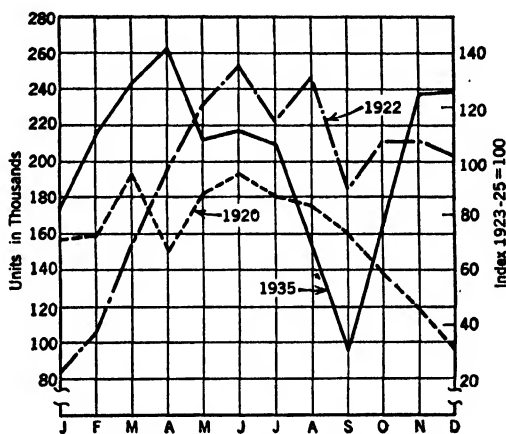

Fig. 12·6.—Changing Seasonals in Automobile Production. Data from 1920 to 1922 in thousands; 1935, relatives on 1923-1925 base. Source: United States Department of Commerce.

[1] It may be said in criticism of this method that the seasonal index itself should first be corrected by removing from the seasonal relatives the effect of the changing Easter. This modification might be desirable but is not essential. In the long run, with adequate data, it would have no appreciable effect, inasmuch as the number of points shifted from April to March would be the same as the number shifted in the opposite direction. It may be added, also, that the correction for Easter may be based upon whatever number of years appears to be reasonably adequate. Or it could be estimated informally by inspection.

is not alone attributable to road building; it reflects also a number of other developments, including a determined effort on the part of producers to change buying habits to regularize production. This shift in automobile production and sales is reflected in a similar readjustment in the sales of petroleum products, for the year-round use of cars has tended to reduce the peak of sales in this related. industry.

In other instances, legislation may directly affect seasonal tendencies in sales, production, and other aspects of business. Thus, the volume of December sales of securities on the stock exchanges of the nation has tended to increase since the enactment of income-tax legislation, for the latter encourages sales to establish losses to be included in income-tax reports as of December 31. Similarly, custom and tradition may change and thus effect extensive changes in the seasonal demands for certain goods, as has been the case with low shoes or oxfords, formerly worn only in the summer. Again, the development of new products and techniques may permit or encourage such shifts. In the building industry, for instance, it was formerly believed necessary to cease construction with the first frost and to renew operations only when the frost "went out" in the spring. With improvements in heating facilities and the development of quick-setting cement, together with numerous related and similar changes, it has been found possible and profitable to continue a great deal of such construction throughout the year. Naturally, seasonal fluctuations in industries dependent upon construction have been affected by this shift.

Measuring the change in seasonals.—In department-store sales, analyzed in preceding paragraphs, similar changes may occur. For example, the Christmas trade may gradually come to occupy a greater or less proportion of the year's business. Under such circumstances, an index that is an average of a period of years may fail to adjust the data satisfactorily, particularly in the very early and late years.

The tendency of seasonals to change may generally be discovered by an inspection of the table of seasonal relatives or by

a line chart of these relatives.[1]   It is obvious that, if a change in the seasonal is taking place, an increase in the business of any given month must be offset by declines in other months, since in any year the average of the seasonal relatives approximates 100.   Of course, over a short period of years, a purely accidental shifting may be taking place.   But if the change seems to be fairly marked, it may be desirable to take account of it.

The simplest way of allowing for changing seasonals is by the computation of successive indexes, covering only a few years.   For example, assuming more complete data than that of Example 12·1, the seasonal index for 1925 might be obtained from the seasonal relatives of 1924, 1925, and 1926, and that for 1926 from the seasonal relatives of 1925, 1926, and 1927, and so on.   By this method a continuously changing index can be obtained.   For limited or current data, the first and last indexes can be estimated by graphic extrapolation, or the contiguous index as calculated could be repeated.   The criticism that erratic items may thus be given too much importance may be remedied by extending the span of years covered by each index to five or seven years.

· Another possible method, applicable to fairly regular data, is that of fitting trends to the seasonal relatives.   The main objection to this method is that it is too inflexible and does not readily adapt itself to the irregularities of change.   However, to illustrate the method, it has been experimentally applied to the data of Example 12·1, and the data corrected for seasonal influences thus measured are plotted in Fig. 12·7, together with the data as previously corrected, and the moving average. Since this method is not a common procedure, it is not illustrated here.

In cases of extremely irregular seasonals, it may be necessary to abandon the attempt to express any average or trended seasonal index.   Under such circumstances, the data adjusted

---

[1] The successive January relatives may be plotted on one chart, the February relatives on another, and so on.   Or, the link relatives (ratios to preceding month; see below, Example 12·4) may be similarly plotted.   (See E. C. Bratt, *Business Cycles and Forecasting*, Business Publications, Inc., Chicago, 1940, p. 26.)

for seasonal are simply the items of the moving average. As applied to historic data, this method does not call for the seasonal relatives at all. But as applied to current data, the seasonal relatives are required as a basis for a tentative current seasonal index.

The adjustments already illustrated by no means exhaust the many possibilities of refining the measurement of seasonal variability. In the case of retail sales, it might be possible to
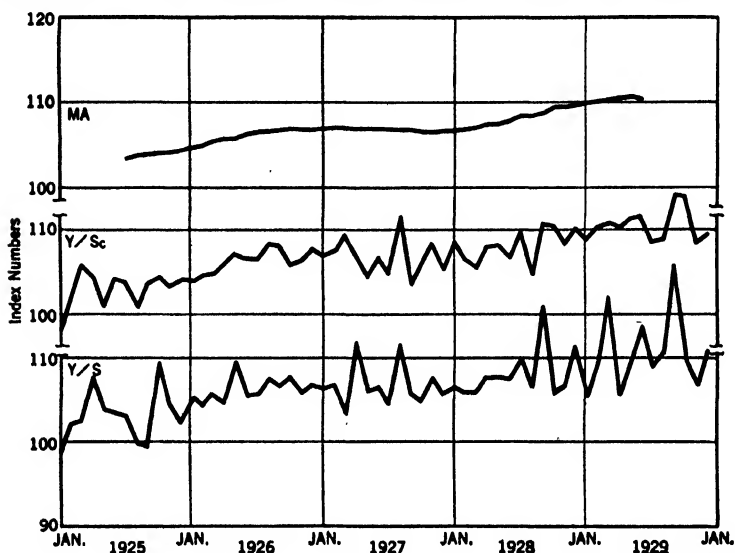

Fig. 12·7.—Department-Store Sales, 1925–1929, Corrected for Seasonality, *MA*, 12-months moving average, centered; Y/S_o, data corrected for changing (trend and Easter) seasonal; Y/S, data corrected for average seasonal. (See Examples 12·1 and 12·2, and page 293.

obtain data respecting weather and to make allowances for an early or a late spring and an early or late fall, somewhat after the fashion of the Easter correction. However, this adjustment is probably not practical at the present time. An example of a really practical adjustment, however, arises in cases like that of automobile sales where, as previously noted, the seasonal is materially affected by the date of introduction of new models. For example, in 1935 when new models were introduced somewhat earlier than usual, the seasonal was shifted accordingly.

The Federal Reserve Board, in working out its index of adjusted industrial production, made a study of this factor in earlier years and corrected its index accordingly. In 1937, 1938, and 1939, similar adjustments were necessary.

A compromise method which has been utilized with some success abandons the idea of an annual seasonal index. The seasonal relatives (Y/MA) are computed as before. Then each Y is corrected by dividing it by an average of the seasonal relatives for the same month in three or four preceding years. For example, the Y of June, 1940, might be divided by an average of the June seasonal relations of 1937, 1938, and 1939, weighted successively by 2, 3, and 5. Or the average might be extended over more years, with weightings stressing the later years, as required. Such a procedure is a compromise, but it may prove practical in some cases.

Flexibility of seasonal indexes.—Seasonal indexes to be used as adjustment factors for raw data may vary widely between two extremes (see Fig. 12·7). At one extreme, there is the average index obtained from the data of many years and applied uniformly, in which case variations in the adjusted data may be individually explained according to circumstances. At the other extreme, such seasonal indexes may be highly flexible, changing from year to year with all the variable factors that are measurable, or they may be merely the seasonal relatives themselves, in which event the moving average is regarded as the data corrected for seasonal. In actual practice, the extent of this flexibility is a matter of judgment.

The link-relative method.—Seasonal indexes are sometimes calculated by a method which differs radically from the moving-average method, as just described. This method is known as the *link-relative* method.[1] The link-relative method has the advantage of being somewhat shorter, but it is not so convenient for purposes of detailed adjustment as the moving-average method, nor is it generally considered quite so accurate. It begins with the calculation of the so-called link relatives, or

---

[1] As an illustration of its use, see indexes of sales of general merchandise in small towns and rural areas, prepared by the Marketing Research Division of the United States Department of Commerce.

ratios of each item (except the first) to the preceding, expressed
as a percentage. The procedure is as follows (see Example 12·4):
(1) Linking: Divide the second item of the given time

EXAMPLE 12·4

SEASONAL INDEXES, LINK-RELATIVE METHOD

Data: Index of retail sales, see Example 12·1.

Month	1925	1926	1927	1928	1929	M of Md items
January		51.1	49.5	50.0	46.9	49.75
February	101.2	96.7	97.8	96.7	101.1	98.53
March	110.6	111.5	106.7	110.2	117.6	110.77
April	111.7	105.2	114.7	108.2	96.3	108.37
May	98.1	ˏ106.9	96.3	101.9	105.8	101.93
June	95.1	91.7	96.2	95.3	99.1	95.53
July	76.5	77.0	75.2	78.4	73.1	76.23
August	101.3	106.5	111.8	101.2	106.3	104.70
September	127.6	126.8	121.2	139.5	139.3	131.23
October	125.8	115.4	113.6	104.4	104.3	111.13
November	100.0	103.3	107.7	105.9	102.5	103.90
December	144.3	148.4	144.4	153.6	152.8	148.53

Month	Chain	Trend Correction	Crude index	Final index
January	100.00	0.0	100	84.1
February	98.53	−0.24	98.29	82.7
March	109.14	−0.49	108.65	91.4
April	118.28	−0.73	117.55	98.9
May	120.56	−0.98	119.58	100.6
June	115.17	−1.22.	113.95	95.9
July	87.79	−1.46	86.33	72.6
August	91.92	−1.71	90.21	75.9
September	120.63	−1.95	118.68	99.9
October	134.06	−2.20	131.86	111.0
November	139.29	−2.44	136.85	115.2
December	206.89	−2.68	204.21	171.8
January (projected)	102.93	−2.93	12)1,426.16	1,200.0
			118.84$\frac{2}{3}$	

series by the first item to obtain the link relative of the second item; the third by the second, to obtain the link relative of the third; and so on, to the end of the series. Each of the resulting link relatives may be regarded as an index number having the preceding item as a base. They are tabulated like the seasonal percentages, and averaged as follows:

(2) Averaging: Average by like periods (the average of the link relatives for each January, then for each February, etc., or similarly by quarters for quarterly data). The median or the average of a few of the central items in the array is generally used, thus eliminating the effects of extreme and irregular fluctuations. Thus 12 link relatives are obtained, one for each month, expressing the typical month-to-month change.

(3) Chaining: Chain the averages thus obtained, taking the first (January) as 100. Multiply 100 by the second, the product thus obtained by the third, etc., thus reversing the calculation of the link relatives. The last chain item (December) is multiplied by the January link relative to obtain a projected January. If no trend effect is present, the last item in the index will also be 100.

(4) Leveling: If the last item in the chain index thus obtained is above or below 100, subtract (algebraically) from each item the fraction of the discrepancy corresponding to the successive given months, as $\frac{1}{12}$ from February, $\frac{2}{12}$ from March, etc. The projected January will then be 100 in agreement with the first January, and the others will be changed proportionately from the constant 100 of the first January. In general, if the discrepancy is d (d = projected item of the chain less 100), and the number of subdivisions in the year is s, subtract from the successive items of the crude index after the first (100), respectively, $\frac{d}{s}$, $\frac{2d}{s}$, $\frac{3d}{s}$ ... $\frac{sd}{s}$. The result is a leveled index from which trend influence has presumably been removed.[1]

---

[1] Strictly speaking, this leveling of the index should be done by the use of logarithms, inasmuch as it involves a series of multiplications. This may be done by expressing the chain as logarithms and reducing the projected figure to 2 (or to 0 if the chain is written in decimals) in a manner similar to that just described. The antilogarithms are then taken. Ordinarily this variation in method is not significant except when the adjustment is large.

(5) Centering: The crude index without the projected January may now be centered, so that its base is 100, as in the moving-average method. This is done by dividing each item by the average of all the items. The results, expressed as percentages, constitute the final index of seasonal variations (S).

An inspection of the method as illustrated in Example 12·4 will indicate that, although the final steps in adjusting the index are somewhat complex, the method as a whole is considerably shorter than the moving-average method. Instead of the calculation of the moving average and division of the data by it, there is now simply the finding of the link relatives. Hence, the method saves almost all the time required for calculating a moving average.

Still other methods of computing a seasonal index might be cited as short-cut approximations. One of these is the so-called *ratio-to-trend method*, in which the seasonal relatives are computed on the basis of an appropriate trend instead of the moving average. Except for this variation the method is the same as the moving-average method already described. Another is the *simple average method*, which involves averaging the data by months or other seasonal intervals to obtain a composite index (the averages generally excluding extreme items), calculating a straight-line trend having the same slope as the annual data and the same total as the seasonal index, and expressing the index in percentages of this trend. Both methods may give satisfactory results with regular data. Graphic methods, also, may be utilized.

Reliability of a seasonal index.—It is often desirable to test seasonal data to determine whether the pattern by months is relatively consistent, that is, whether it recurs more often than might be expected merely by chance fluctuation. The basis on which such a test is made will be discussed later (Chapter XIX), but the procedure may be outlined here. In the first place, the seasonal relatives are tabulated by years (beginning with the month of the first seasonal relative, usually July; see Example 12·5), and are ranked within each such year consecutively from lowest to highest. These rankings are then added for each month, and the sums labeled X. It is obvious that the vari-

ability of X is a measure of the persistence of the seasonal pattern. Hence, in effect the variance, $\Sigma x^2 / N$, is computed. This variance may be expressed as a ratio to the mean, or, more conveniently, as $\Sigma x^2 / \Sigma X$. If this ratio is multiplied by 6, its likelihood of recurring by mere chance may be evaluated by

EXAMPLE 12·5

RELIABILITY OF A SEASONAL INDEX

Data: Monthly seasonal relatives of retail sales, July, 1925, to June, 1929 (see Example 12·1, page 278), expressed as ranks (smallest to largest in each fiscal year).

Month	Fiscal years				Sum of Ranks	
	1925–6	1926–7	1927–8	1928–9	X	X^2
July	(1) 72.6	(1) 72.4	(1) 71.3	(1) 73.9	4	16
August	(2) 73.3	(2) 77.0	(2) 79.8	(2) 74.8	8	64
September	(6) 93.4	(7) 97.7	(7) 96.6	(9) 103.9	29	841
October	(11) 117.4	(10) 112.5	(10) 109.9	(10) 108.1	41	1,681
November	(10) 117.3	(11) 116.1	(11) 118.4	(11) 114.5	43	1,849
December	(12) 168.7	(12) 172.4	(12) 170.9	(12) 175.3	48	2,304
January	(4) 86.1	(4) 85.3	(4) 85.2	(3) 82.0	15	225
February	(3) 83.0	(3) 83.3	(3) 82.4	(4) 82.9	13	169
March	(5) 92.0	(5) 88.9	(5) 90.7	(6) 97.2	21	441
April	(8) 96.6	(9) 102.2	(8) 97.7	(5) 93.3	30	900
May	(9) 103.2	(8) 98.4	(9) 99.5	(8) 98.6	34	1,156
June	(7) 94.3	(6) 94.7	(6) 94.6	(7) 97.7	26	676
					12)312	10,322
					26	8,112

$$\chi_r^2 = \frac{6\Sigma x^2}{\Sigma X} = \frac{6 \times 2,210}{312} = 42.50 \qquad M_x = 26$$

$$\Sigma x^2 = 2,210$$

Least significant $\chi_r^2 = 20$

Least highly significant $\chi_r^2 = 25$

means of a table of chi square (see page 561). The table is read for $N - 1$, or, with monthly data, $12 - 1 = 11$, and indicates a 5 per cent probability of about 20 and a 1 per cent probability of 25. These values may therefore be considered as the least significant and the least highly significant values, respectively.

In Example 12·5, for instance, chi square by ranks is found to be 42.50. The seasonal factor may therefore be considered highly significant.

READINGS

(See also special and general references, pages 591 and 597.)

BAUMAN, A. O., "Thirteen-Months-Ratio-First-Difference Method of Measuring Seasonal Variations," *Journal of the American Statistical Association*, 23 (M.S. 163), September, 1928, pp. 282–290.

BRATT, ELMER C., *Business Cycles and Forecasting*, Chicago, Business Publications, Inc., 1937.

BURSK, J. PARKER, "Seasonal Variations in Employment in Manufacturing Industries—A statistical Study based on Census Data," *University of Pennsylvania, Wharton School of Finance and Commerce, Industrial Research Studies, No. 14*, October, 1931.

COVER, JOHN H.; REVZAN, DAVID A.; HELMS, WILFRED M.; and COHENOUR, VINCENT J., "A Barometer of Chicago Stock Prices," *Journal of Business* (University of Chicago), 3 (2), April, 1930, pp. 170–191.

FOX, BERTRAND, "Seasonal Variations in Selected Series of Weekly Data," *Review of Economic Statistics*, 13 (1), February, 1931, pp. 26–33.

FRICKEY, EDWIN, "A Statistical Study of Bank Clearings, 1875–1914," *Review of Economic Statistics*, 12 (2), May, 1930, pp. 90–99; 12 (3), August, 1930, pp. 112–138.

FRICKEY, EDWIN, "Outside Bank Debits Corrected for Seasonal Variation: Monthly and Weekly, 1919–1931," *Review of Economic Statistics*, 13 (2), May, 1931, pp. 76–84.

HOMAN, JEANETTE, "Adjusting for the Changing Date of Easter in Economic Series," *Journal of the American Statistical Association*, 28 (183), September, 1933, pp. 328–333.

JOY, ARYNESS, and THOMAS, WOODLIEF, "Use of Moving Averages in the Measurement of Seasonal Variations," *Journal of the American Statistical Association*, 23 (163), September, 1928, pp. 241–252.

KUZNETS, SIMON, "On the Analysis of Time Series," *Journal of the American Statistical Association*, 23 (164), December, 1928, pp. 398–410.

—— "Seasonal Pattern and Seasonal Amplitude; Measurement of Their Short-time Variations," *Journal of the American Statistical Association*, 27 (177), March, 1932, pp. 9–20.

MACAULEY, FREDERICK R., "The Smoothing of Time Series," National Bureau of Economic Research, *Monograph No. 19*, February, 1931.

PALMER, EDGAR Z., "Error and Unreliability in Seasonals," *Annals of Mathematical Statistics*, I (4), November, 1930, pp. 345–352.

PISER, LEROY M., "A Method of Calculating Weekly Seasonal Indexes," *Journal of the American Statistical Association*, 27 (179), September, 1932, pp. 307–309.

ROBB, RICHARD A., "Variate Difference Method of Seasonal Variation," *Journal of the American Statistical Association*, 24 (167), September, 1929, pp. 250–257.

SPURR, W. A., "A Graphic Method of Measuring Seasonal Variations," *Journal of the American Statistical Association*, 32 (198), June, 1937.

—— "A Graphic Short Cut to the Moving Average Method of Measuring Seasonality," *Journal of the American Statistical Association*, 35 (212), December, 1940.

SZELISKI, VICTOR S., von, "Comment on 'The Variate Difference Method of Seasonal Variation,'" *Journal of the American Statistical Association*, 25 (169), March, 1930, pp. 83–84.

WELCH, EMMETT H., "Adjusting Indexes of Seasonal Variation for Trend," *Journal of the American Statistical Association*, 25 (172), December, 1930, pp. 440–446. "A Mathematical Theory of Seasonal Indices," *Annals of Mathematical Statistics*, 1 (1), February, 1930, pp. 57–72.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Using the moving-average method, compute a seasonal index from the quarterly data listed below.

    (a) 49, 50, 54, 66; 75, 72, 72, 80; 85, 78, 74, 78.

    (b) 84, 68, 60, 63; 62, 50, 46, 53; 56, 48, 48, 59.

    (c) 94, 93, 88, 72; 64, 67, 66, 54; 50, 57, 60, 52.

    (d) 62, 64, 65, 76; 80, 78, 75, 82; 82, 76, 69, 72.

    (e) 52, 54, 55, 66; 70, 68, 65, 72; 72, 66, 59, 62.

    (f) 90, 89, 84, 68; 60, 63, 62, 50; 46, 53, 56, 48.

    (g) 96, 99, 96, 82; 78, 85, 86, 76; 76, 87, 92, 86.

    (h) 65, 84, 95, 90; 93, 104, 107, 94; 89, 92, 87, 66.

    (i) 85, 90, 87, 71; 67, 76, 77, 65; 65, 78, 83, 75.

    (j) 66, 80, 87, 78; 84, 94, 97, 84; 86, 92, 91, 74.

    (k) 12, 52, 64, 104; 100, 124, 120, 144; 124, 132, 112, 120.

2. Using the moving-average method, compute approximate measures of seasonal variations in the quarterly data summarized below.

    (a) 80, 59, 104, 119; 84, 63, 116, 137; 97, 72, 127, 142.

    (b) 92, 91, 102, 107; 90, 96, 103, 111; 94, 95, 107, 112.

    (c) 88, 68, 104, 126; 87, 70, 113, 132; 94, 72, 114, 138.

    (d) 90, 93, 91, 94; 98, 101, 99, 102; 106, 109, 107, 110.

    (e) 6, 7, 12, 15; 14, 15, 20, 23; 22, 23, 28, 31.

    (f) 90, 96, 168, 176; 162, 160, 264, 264; 234, 224, 360, 352.

    (g) 48, 49, 55, 56; 56, 57, 63, 64; 64, 65, 71, 72.

3. Each of the following series represents quarterly data for three consecutive years. Using the link-relative method, compute approximate measures of seasonal variations.

    (a) 80, 59, 104, 119; 84, 63, 116, 137; 97, 72, 127, 142.

    (b) 88.4, 92, 94.6, 93; 95.9, 100, 103.1, 101; 104, 108.1, 111.2, 108.7

    (c) 209, 205, 211, 207; 201, 197, 203, 199; 193, 189, 195, 191.

    (d) 208, 204, 210, 206; 200, 196, 202, 198; 192, 188, 194, 190.

    (e) 192, 196, 190, 194; 200, 204, 198, 202; 208, 212, 206, 210.

4. From the following quarterly seasonal percentages $(Y \div MA)$ for the years 1937–1940, calculate centered indexes of seasonal variation:

(a) Seasonal percentages.

Quarter	1937	1938	1939	1940
1	...	40	80	60
2	...	100	120	80
3	80	60	100	..
4	120	140	160	..

(b) Seasonal percentages.

Quarter	1937	1938	1939	1940
1	...	70	60	80
2	...	80	100	90
3	140	130	150	..
4	110	130	120	..

ANSWERS TO EXERCISES

1. Centered MA, SR to 1 decimal.

(a) 107.2; 97.8; 93.1; 101.9.   (g) 93.7; 104.6; 106.8; 94.9.
(b) 110.7; 93.6; 90.8; 104.9.   (h) 94.2; 103.7; 106 9; 95 2.
(c) 91.6; 104.7; 108.5; 95.2.   (i) 91.4; 106 7; 109.1; 92 8.
(d) 104.2; 99.1; 93.7; 103.0.   (j) 95.7; 104.7; 107.3; 92.3.
(e) 104.9; 98.9; 92.7; 103.5.   (k) 93.3; 103.2; 92.1; 111.4.
(f) 91.0; 105.1; 109.1; 94.8.

2. Centered MA, SR to 1 decimal.

(a) 90.1; 65.3; 114.3; 130.3.   (e) 93.5; 89.1; 106.1; 111.3.
(b) 92.0; 94.9; 103.5; 109.6.   (f) 90.6; 79.5; 119.0; 110.9.
(c) 90.3; 69.9; 109.9; 129.9.   (g) 98.3; 96.7; 103.4; 101.6.
(d) 101.1; 102.0; 97.9; 99.0.

3. (a) 89.3; 64.6; 115.1; 131.0.   (d) 99.0; 98 0; 102.0; 101.0.
(b) 99.0; 101.1; 102.0; 97.9.   (e) 101.0; 102.0; 98.0; 99.0.
(c) 99.0; 98.1; 101.9; 101.0.

4. (a) 63, 105, 84, 148.   (b) 67, 86, 133, 114.

B. PROBLEMS

5. Making use of any of the accepted methods described in the chapter, calculate a seasonal index for the following combined quarterly index of industrial production.

Source: *Survey of Current Business Year Book*, 1932.

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	102.00	102.00	106.33	107.67	110.00	109.33	120.67	106.00
2	106.67	90.00	102.33	107.00	109.67	109.33	125.00	103.67
3	100.67	87.67	100.67	108.33	104.33	110.33	121.67	91.00
4	97.67	98.00	106.00	108.33	100.67	114.00	108.33	83.67

6. Prepare an index of seasonal variations for the following data of freight-car loadings as released by the Federal Reserve Board:

Quarter	1923	1924	1925	1926	1927	1928	1929	1930
1	90.67	93.33	94.67	96.33	99.00	94.33	97.33	90.00
2	100.67	92.67	100.33	104.33	103.00	100.67	107.00	95.00
3	107.33	101.33	109.67	114.33	109.67	111.00	115.67	96.67
4	100.67	103.00	106.33	111.00	101.00	107.33	103.00	85.67

7. Use the data below, representing actual monthly payrolls for three specified firms, to calculate: (1) seasonal relatives by the ratio to moving averages method (use a 12 months' moving average centered in the seventh month); (2) seasonal relatives by the link-relative method.

FIRM A: FOOD PROCESSING

	1926	1927	1928	1929	1930	1931
January	2931	2779	2720	2510	2768	2786
February	2594	2158	2514	2329	2636	2598
March	2428	2299	2405	2392	2426	2394
April	2351	2327	2395	2343	2337	2406
May	2350	2368	2411	2349	2397	2499
June	2570	2428	2454	2491	2595	2482
July	2599	2300	2340	2554	2482	2481
August	2405	2358	2244	2417	2438	2463
September	2594	2327	2417	2536	2621	2425
October	2871	2768	2778	2864	2938	2657
November	3134	2925	2913	2966	2858	2658
December	2883	2764	2685	2604	2482	2658

FIRM B: METAL MANUFACTURE

	1926	1927	1928	1929	1930	1931	1932
January	1141	1123	1124	1216	1069	894	487
February	1180	1172	1162	1255	1186	899	536
March	1237	1188	1217	1298	1306	872	530
April	1303	1278	1291	1341	1328	828	495
May	1353	1293	1351	1377	1206	886	478
June	1337	1305	1364	1370	1103	668	442
July	1313	1257	1325	1343	1103	596	439
August	1261	1214	1297	1316	1022	552	420
September	1196	1167	1257	1005	973	543	407
October	1133	1140	1210	980	881	537	392
November	1082	1108	1207	963	854	452	367
December	1102	1103	1201	990	848	372	342

FIRM C: METAL MANUFACTURE

	1926	1927	1928	1929	1930	1931
January	322	314	340	1045	564	646
February	343	345	429	1103	688	633
March	368	358	439	1129	771	639
April	411	392	506	1186	820	626
May	453	442	579	1238	826	641
June	512	489	653	1371	657	634
July	544	523	703	1259	742	632
August	548	536	787	1218	847	687
September	496	537	880	1207	833	783
October	472	461	944	1123	725	824
November	388	407	1021	1084	758	754
December	343	343	739	643	700	581

8. The following tabulation shows the average farm price of eggs, by months, 1932–1935, as adapted from *Statistical Abstract of the United States*, 1935. Calculate a seasonal index by the method of link relatives, and correct the data for seasonal.

	1932	1933	1934	1935
January	17	21	18	25
February	13	11	16	26
March	10	10	14	19
April	10	10	14	20
May	10	12	13	21
June	11	10	13	21
July	12	13	14	22
August	15	13	17	23
September	17	16	22	26
October	22	21	24	27
November	26	24	29	29
December	28	22	27	28

9. Given the following median link relatives; compute an index of seasonal variation.

MONTH	MEDIAN LINK RELATIVES	MONTH	MEDIAN LINK RELATIVES
January	90	July	80
February	80	August	120
March	100	September	110
April	110	October	100
May	90	November	80
June	100	December	90

10. Given the following series of bank deposits in hundreds of thousands of dollars, and the seasonal index in relative form, adjust the original series for seasonal variation.

If the trend were practically horizontal, what conclusion with respect to the cyclical position of the series would be justified? Write one sentence describing the course of the cycle over the 6-month period. How would you adjust the data if the seasonal were stated *in absolute form?*

	MONTH	DEPOSITS	SEASONAL
1935	January	74	90
	February	63	80
	March	80	82
	April	94	82
	May	106	81
	June	139	102

11. Recalculate the seasonal index of Example 12·1, page 278, taking the moving average on a direct 12-month basis, centered in the seventh month.

12. Consult the *Survey of Current Business* for further data, and prepare indexes of seasonal variations.

CHAPTER XIII

CYCLICAL VARIATIONS

Preceding chapters devoted to the analysis of time series have first described the secular trend or "regular and persistent change in a variable during a long period of time" and have explained how it may be measured. They have then given attention to methods of discovering and measuring seasonal patterns in such data. In each case, consideration has been given also to means by which data may be adjusted to take account of or "correct" for the discovered tendencies. When data have been adjusted for trend and seasonality, the remaining variability is generally described as representing two types of change, the one *cyclical* and the other *episodic* or *residual* (random). *Cyclical fluctuations* are those whose more or less regular rise and fall mark the various phases of business prosperity and depression.[1] *Episodic fluctuations* are those which reflect a single major event, such as a war, a flood, or a panic. *Residual* or *random fluctuations* are those which remain after data have been adjusted for trend, seasonality, and cyclical variation. In this chapter, attention is directed to methods of discovering and measuring cyclical fluctuations, with which episodic and random changes are usually included.

As has been noted in Chapter XI, when data representing a considerable period of time are analyzed, a fairly well-defined secular trend may often be noted. If it is considered as repre-

---

[1] Students of the business cycle have made a case for three types of cycles, a long wave covering about half a century, an intermediate cycle of 8 or 9 years, and the familiar cycle of 3 or 4 years. Major depressions are accounted for, according to this theory, by the concurrence of two or more of these cycles. The types are known by the names of men who investigated them, namely, Kondratieff, Juglar, and Kitchen, respectively.

sentative of *normal*, i.e., the general growth tendency of the data, the ratio of any individual item in the series to the trend value for the same date will indicate how closely the actual data approximate normal, and the difference between such a ratio and 1 is a measure of deviation from normal and thus of cyclical and residual fluctuation. Such a ratio is called a *percentage cycle* (symbol $C\%$). Where annual data are used, the ratio may be calculated directly. If weekly, monthly, or quarterly data are used, however, the seasonal pattern must first be removed by dividing actual data by the seasonal index, to secure Y/S, as has been indicated in Chapter XII. A suitable trend value computed first from the annual data and then adjusted to match each weekly, monthly, or quarterly item, as described later, may then be used as a second divisor. The result may be described as Y/ST, i.e., data adjusted for seasonal and trend. Or the same result may be obtained by multiplying the trend by the seasonal index, S, and regarding the product as the normal, with which Y may then be compared. Occasionally comparisons are made by subtracting rather than dividing, but in general ratios are more suitable, since they tend to equalize average deviations.

As has been noted, in Chapters X and XI, there are many cases in which it is not an easy matter to select a suitable trend. For a comparatively few years, a straight line may appear quite satisfactory for most data, while, over longer periods, population and production may typify a growth curve, and prices over considerable periods of time frequently conform to the shape of a parabola. But assuming that a suitable trend has been fitted to a broad sample of production data, the percentages obtained from the ratio, Y/T, or, with seasonal items, Y/ST, represent, broadly speaking, a measure of influence of the business cycle, although it must be recognized that the term "cycle" is not to be interpreted strictly, since actual business cycles vary greatly in length and may reflect random change, as well as the effects of accidental or *episodic* change, such as drought and war. Hence, the term "business cycle" may be misunderstood if it is regarded as implying a certain regularity of wavelike ups and downs of business. The general

business cycle, for instance, when charted over a long period of time, appears more like a chance grouping of accidental happenings.[1]

Historical vs. current studies.—It should also be emphasized that the measurement of the business cycle belongs more strictly to historic studies than to current interpretation. Prosperity and depression are, after all, relative terms. After several comparable cycles have been completed, it is quite possible to measure each by reference to a trend line fitted to the data and regarded as representing *normal* for the entire period. And as long as cyclical change holds within fairly narrow limits along such a trend, as it did for more than a century before 1929, it is possible to estimate plausibly the current phase of the cycle. The depression since 1929, however, has broken all precedents. Even before political measures modified economic processes, it had registered, in percentage terms, roughly twice as deep a trough as ever before. Subsequent advances and relapses can therefore not be appraised until time has given additional evidence to the long-time secular trend.

As a result of uncertainty concerning the current secular trend, reference is sometimes made to an evaluation of the current position in terms of a pre-depression base, such as 1923–1925, instead of to a fitted trend. The *Business Week* index, revised in 1938, may be cited as evidence of this practice. Other indexes, including the former *Annalist* index of business, make use of trends based on an inclusion of depression data, thus tending to elevate the current phase of the cycle. *Barron's* index, on the other hand, states the current position both as a percentage of a projected normal (based largely on pre-depression data) and as a percentage of a fixed base. The Cleveland Trust Company estimate of the cycle also employs a projected trend (see Fig. 13·1). Obviously the bases of measurement should

[1] The length of cycles, up to the time of the great depression, for all the larger countries, and for the United States taken separately, conforms rather strikingly to the logarithmic normal curve, which is an expression of certain laws of chance. See W. C. Mitchell, *Business Cycles, The Problem and Its Setting*, National Bureau of Economic Research, 1927, New York, p. 419, and G. R. Davies, "The Analysis of Frequency Distributions," *Journal of the American Statistical Association*, December, 1929, pp. 365–366.

be taken into account by those who would interpret such indexes of current business activity.

Measurement of cycles in annual data.—The measurement of the cycles (in percentages of trend or normal) for annual data
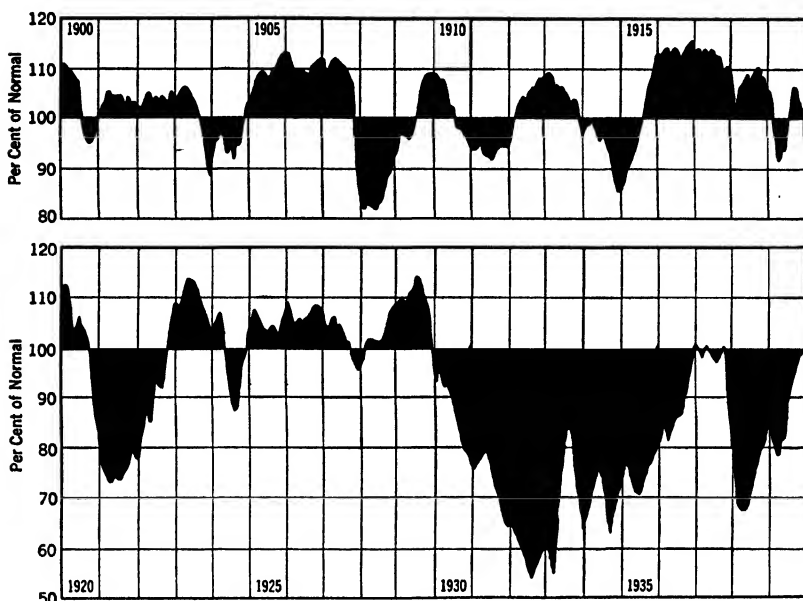

Fig. 13·1.—Business Cycles in the United States, 1900–1939. Business activity in percentages of a computed normal. Adapted from the Cleveland Trust Company *Business Bulletin,* by permission.

is illustrated in Example 13·1. The data used there represent annual indexes of industrial production for the years 1910–1929 inclusive. For the years from 1919 through 1929, the data have been computed by the statistical division of the Federal Reserve Board; the data for the earlier years are estimates based on calculations made by Col. L. P. Ayres.[1]

A linear trend has been fitted to the data and is itemized in the example.[2] It is defined by the equation

$$T = 86.8\underline{6} + 2.43x \text{ (Origin at Jan. 1, 1920)}$$

---

[1] See his *Turning Points in Business Cycles*, New York, the Macmillan Co., 1940, p. 203.

[2] A preliminary trend was fitted by the method of grouped data. Then, because one or two extremely low deviations appeared, the trend was centered to set it in a

EXAMPLE 13·1

THE PERCENTAGE CYCLE—ANNUAL DATA

Data: Industrial production in the United States, 1910–1929. Estimated, 1910–1918; Federal Reserve Board, old series, 1919–1929. Trend is $T = 86.\underline{86} + 2.43x$ (underscored digit is a repetend); origin at January 1, 1920.

Year	Y Midyear	Trend T	Per cent cycle Y/T	Per cent deviation $d\%$	AD cycle $d\%/AD$
1910	62	63.8	97.2	−1.635	−0.27
1911	60	66.2	90.6	−8.235	−1.37
1912	69	68.6	100.6	1.765	0.29
1913	71	71.1	99.9	1.065	0.18
1914	65	73.5	88.4	−10.435	−1.73
1915	72	75.9	94.9	−3.935	−0.65
1916	86	78.4	109.7	10.865	1.80
1917	87	80.8	107.7	8.865	1.47
1918	85	83.2	102.2	3.365	0.56
1919	83	85.7	96.8	−2.035	−0.34
1920	87	88.1	98.8	−0.035	−0.01
1921	67	90.5	74.0	−24.835	−4.12
1922	85	92.9	91.5	−7.335	−1.22
1923	101	95.4	105.9	7.065	1.17
1924	95	97.8	97.1	−1.735	−0.29
1925	104	100.2	103.8	4.965	0.82
1926	108	102.7	105.2	6.365	1.06
1927	106	105.1	100.9	2.065	0.34
1928	111	107.5	103.3	4.465	0.74
1929	119	110.0	108.2	9.365	1.56
	1,723	1,737.4	1,976.7	−60.215	−10.00
			$M = 98.835$	+60.215	9.99
				20)120.430	19.99
				$AD = 6.0215$	

median position. The procedure utilized for this purpose consists of (1) finding deviations from the preliminary trend $(Y - T_p)$, (2) ranking these deviations, and (3) averaging the median items. The seven largest and smallest items were discarded in this case. The average of the remaining six (0.71$\underline{6}$) was added to $a = 86.15$, as first computed, to obtain the revised $a = 86.8\underline{6}$. As thus revised, the trend supplies more plausible projections of the normal to later dates.

The cycle in percentages of trend (the percentage cycle) was then obtained as Y/T. These values are written as percentages, and the deviations from the average of these values were calculated and plotted with the data and trend in Fig. 13·2. These deviations measure cyclic change together with minor random and episodic fluctuations.

For some purposes it is useful to express the cycle in terms of average or standard deviations, hence the so-called *AD cycle* has also been calculated as

$$AD \text{ cycle} = \frac{d\%}{AD}$$

Theoretically, the absolute sum of these items should equal N, and their algebraic sum should be 0. Their utility lies partly in the fact that they more clearly indicate extreme deviations, and they are useful as means of comparing and combining various series.

The cycle projected.—In spite of the fact that the depression since 1929 has rendered the secular trend extremely uncertain, this trend as defined by pre-depression years is often extrapolated. Such extrapolation assumes that technical progress and population growth have continued to advance potential productive capacity as measured by the computed trend. In other words, it assumes that the balance between depression and later recovery will approximate the trend which has been established for a long time. The assumption may, of course, be wrong, but it is at least interesting to follow the procedure of several statistical reporting agencies and make the appropriate estimate for recent years. If, for example, the trend as calculated in Example 13·1 is projected to January, 1935, fifteen years and one-half month later than the origin (January 15, 1935, less January 1, 1920), then

$$T = 86.86 + (2.43 \times 15.0416) = 123.42$$

In that month the actual index of industrial production, corrected for the seasonal factor, stood at 90. The measurement of the cycle, therefore, in percentage terms would be

$$C\% = 90 \div 123.4 = 72.9 \ (\%)$$

In other words, in accordance with this estimate, business at the beginning of 1935 was 27.1 per cent below what might theoretically be expected as normal.[1] Whether such an estimate is justifiable or not, two or three widely used indexes are computed on a somewhat similar basis.[2]

TABLE 13·1

INDUSTRIAL PRODUCTION IN THE UNITED STATES, 1923–1930; 1923–25 = 100

(Adjusted for seasonal variations)

Source: Federal Reserve Board, old series.

	1923	1924	1925	1926	1927	1928	1929	1930
January	99	100	105	106	107	107	119	106
February	100	102	104	105	108	108	119	107
March	103	100	103	106	110	108	119	104
April	106	95	102	107	108	108	121	104
May	106	89	102	106	109	108	122	102
June	106	85	102	108	107	108	125	98
July	104	84	103	108	106	109	124	93
August	103	89	103	110	106	110	121	90
September	101	94	101	111	104	113	121	90
October	99	95	104	111	102	115	118	88
November	98	97	107	110	101	117	110	86
December	97	101	109	107	102	118	103	84
Annual	101	95	104	108	106	111	119	96

Cycles in monthly data.—The measurement of cyclical change in monthly data is not essentially different from that just described.   It differs only in respect to the seasonal element and in minor details relating to the trend.   The procedure is illustrated by use of the data of industrial production in the

[1] The Cleveland Trust Company Index of Industrial Production for the same month, adjusted for seasonal and trend, stood at 73.0, or, as recently revised, at 75.7.

[2] The secular trend of industrial production, and the cycle based on the projection of this trend, have been computed from the index of industrial production as published prior to 1940 rather than from the 1940 revision as summarized on page 187. The earlier index was used because revised estimates are not available for years prior to 1923.   However, if the 1923–1929 annual indexes now available were taken as the normal trend base, and the method here described applied, the cycle measure ($C\%$) would not differ much from that just computed.

United States for the eight years from 1923 through 1930, as summarized in Table 13·1. The calculations and results are presented in Tables 13·2, 13·3, 13·4, and 13·5.

TABLE 13·2

TREND OF INDUSTRIAL PRODUCTION, 1923–1930, BY MONTHS

Data: See Example 13·1

	1923	1924	1925	1926	1927	1928	1929	1930
January	94.3	96.7	99.1	101.5	104.0	106.4	108.8	111.3
February	94.5	96.9	99.3	101.8	104.2	106.6	109.0	111.5
March	94.7	97.1	99.5	102.0	104.4	106.8	109.2	111.7
April	94.9	97.3	99.7	102.2	104.6	107.0	109.4	111.9
May	95.1	97.5	99.9	102.4	104.8	107.2	109.6	112.1
June	95.3	97.7	100.1	102.6	105.0	107.4	109.9	112.3
July	95.5	97.9	100.3	102.8	105.2	107.6	110.1	112.5
August	95.7	98.1	100.5	103.0	105.4	107.8	110.3	112.7
September	95.9	98.3	100.7	103.2	105.6	108.0	110.5	112.9
October	96.1	98.5	100.9	103.4	105.8	108.2	110.7	113.1
November	96.3	98.7	101.1	103.6	106.0	108.4	110.9	113.3
December	96.5	98.9	101.3	103.8	106.2	108.6	111.1	113.5
Total	1,144.8	1,173.6	1,202.4	1,232.3	1,261.2	1,290.0	1,319.5	1,348.8

TABLE 13·3

MONTHLY PERCENTAGE CYCLE, $(Y/S) \div T$

Data: See Tables 13·1 and 13·2.

	1923	1924	1925	1926	1927	1928	1929	1930
January	105.0	103.4	106.0	104.4	102.9	100.6	109.4	95.2
February	105.8	105.3	104.7	103.1	103.6	101.3	109.2	96.0
March	108.8	103.0	103.5	103.9	105.4	101.1	109.0	93.1
April	111.7	97.6	102.3	104.7	103.3	100.9	110.6	92.9
May	111.5	91.3	102.1	103.5	104.0	100.7	111.3	91.0
June	111.2	87.0	101.9	105.3	101.9	100.6	113.7	87.3
July	108.9	85.8	102.7	105.1	100.8	101.3	112.6	82.7
August	107.6	90.7	102.5	106.8	100.6	102.0	109.7	79.8
September	105.3	95.6	100.3	107.6	98.5	104.6	109.5	79.7
October	103.0	96.4	103.1	107.4	96.4	106.3	106.6	77.8
November	101.8	98.3	105.8	106.2	95.3	107.9	99.2	76.0
December	100.5	102.1	107.6	103.1	96.0	108.7	92.7	74.0
Total	1,281.1	1,156.5	1,242.5	1,261.1	1,208.7	1,236.0	1,293.5	1,025.5

TABLE 13·4

MONTHLY AVERAGE DEVIATION CYCLE $(d\% = C\% - 100\%)$

Data: See Table 13·3.

	1923	1924	1925	1926	1927	1928	1929	1930
January	+ 5.0	+ 3.4	+6.0	+4.4	+2.9	+0.6	+ 9.4	− 4.8
February	+ 5.8	+ 5.3	+4.7	+3.1	+3.6	+1.3	+ 9.2	− 4.0
March	+ 8.8	+ 3.0	+3.5	+3.9	+5.4	+1.1	+ 9.0	− 6.9
April	+11.7	− 2.4	+2.3	+4.7	+3.3	+0.9	+10.6	− 7.1
May	+11.5	− 8.7	+2.1	+3.5	+4.0	+0.7	+11.3	− 9.0
June	+11.2	−13.0	+1.9	+5.3	+1.9	+0.6	+13.7	−12.7
July	+ 8.9	−14.2	+2.7	+5.1	+0.8	+1.3	+12.6	−17.3
August	+ 7.6	− 9.3	+2.5	+6.8	+0.6	+2.0	+ 9.7	−20.2
September	+ 5.3	− 4.4	+0.3	+7.6	−1.5	+4.6	+ 9.5	−20.3
October	+ 3.0	− 3.6	+3.1	+7.4	−3.6	+6.3	+ 6.6	−22.2
November	+ 1.8	− 1.7	+5.8	+6.2	−4.7	+7.9	− 0.8	−24.0
December	+ 0.5	+ 2.1	+7.6	+3.1	−4.0	+8.7	− 7.3	−26.0

TABLE 13.5

AVERAGE DEVIATION CYCLE $(d\%/AD)$

Data: See Table 13·3.

	1923	1924	1925	1926	1927	1928	1929	1930
January	+0.83	+0.56	+1.00	+0.73	+0.48	+0.10	+1.56	−0.80
February	+0.96	+0.88	+0.78	+0.51	+0.60	+0.22	+1.53	−0.66
March	+1.46	+0.50	+0.58	+0.65	+0.90	+0.18	+1.49	−1.15
April	+1.94	−0.40	+0.38	+0.78	+0.55	+0.15	+1.76	−1.18
May	+1.91	−1.44	+0.35	+0.58	+0.66	+0.12	+1.88	−1.49
June	+1.86	−2.16	+0.32	+0.88	+0.32	+0.10	+2.28	−2.11
July	+1.48	−2.36	+0.45	+0.85	+0.13	+0.22	+2.09	−2.87
August	+1.26	−1.54	+0.42	+1.13	+0.10	+0.33	+1.61	−3.35
September	+0.88	−0.73	+0.05	+1.26	−0.25	+0.76	+1.58	−3.37
October	+0.50	−0.60	+0.51	+1.23	−0.60	+1.05	+1.10	−3.69
November	+0.30	−0.28	+0.96	+1.03	−0.78	+1.31	−0.13	−3.99
December	+0.08	+0.35	+1.26	+0.51	−0.66	+1.44	−1.21	−4.32

In the computation of the cycle for the years just indicated, trend items for each month were obtained by reference to the equation already described as

$$T = 86.8\underline{6} + 2.43x \text{ (Origin, Jan. 1, 1920)}$$

The trend value for January 1, 1923 ($x = 3$), is

$$T = 86.8\underline{6} + (2.43 \times 3) = 94.15\underline{6}$$

The monthly value of the trend slope, b_m, is then found as $b \div 12 = 2.43 \div 12 = 0.2025$. To center the first month's trend value at January 15, one-half this monthly increment or 0.10125 is added to the trend item for January 1, 1923. The trend value for this month is thus $94.15\underline{6} + 0.10125 = 94.3$. Values for subsequent months are calculated by adding successively the slope for one month, $b_m = 0.2025$. In these calculations five decimal places have been carried in order to avoid cumulative errors, but each item when written in the table is rounded to one decimal. These values are summarized in Table 13·2.

· As in the case of annual data, percentage cycle items (the cycle as a percentage of trend) are obtained by dividing the data (in this case Y/S, i.e., seasonally corrected items) by corresponding trend items (T). The percentages thus obtained measure combined cyclic and random fluctuations about the trend, after seasonal and trend influence have been removed.

Again, as in the case of annual data, the cycle may for certain purposes be reduced to units of the average deviation. The objectives in such scaling are, as before noted, the detection of extreme deviations and comparisons with cycles in other data similarly measured. By this means the amplitudes of the cycles are equalized. Results of this procedure are summarized in Tables 13·4 and 13·5. It will be noted that the average deviation is taken from the calculation in Example 13·1, which is assumed to represent a relatively normal period. During the two decades there studied, the average deviation was found to be 6.0215 (in percentages).

As charted in Fig. 13·3, the cycle during the years 1923–1930 reveals the relative prosperity of 1923, the sharp depression following fear of inflation in 1924, the balanced prosperity of 1925–1926, the so-called Ford depression of late 1927 (the delayed debut of Model A), recovery under the shadow of an inflated stock market in 1928, the brink of the depression in late 1929, and the plunge of 1930.
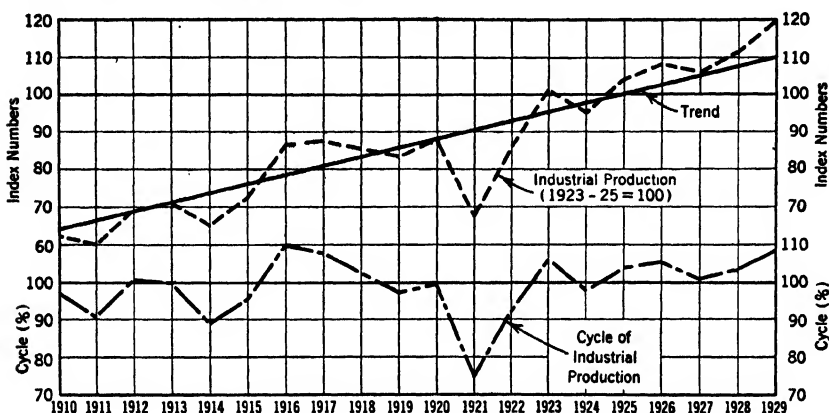

FIG. 13·2.—Annual Index Numbers of Industrial Production, 1910–1929, together with Fitted Trend and Computed Percentage Cycle. (See Example 13·1.) (The 1940 revision of this series is not used because it extends back only to 1923.)

Cycles with complex trends.—When a parabola or other curved trend is indicated, the measurement of the cycle as just described requires modification. If annual data are employed, the trend may be computed as described in Chapter XI, and the percentage cycle found as usual as the ratio of data to trend (Y/T). If the average deviation cycle is required, it may be computed as before. In very long series, the process may be abbreviated by averaging the data in five-year or other convenient intervals, fitting the trend to these averages, and interpolating annual trend items as described below for monthly data. A chart of the data will reveal whether such approximations are reasonably accurate.

With monthly data, it is generally sufficient to calculate the trend from annual data as of the middle of the year, and to interpolate monthly trend items on a linear basis. For example,

suppose that the parabolic trend item for 1926 (as of July 1) is 88.32 and that for 1927 is 89.52. The July trend point (July 15) is then 88.32 increased by one-twenty-fourth of the succeeding annual rise, $89.52 - 88.32 = 1.20$, and the August trend point is further increased by one-twelfth of the annual rise. The required trend points (mid-July, August, etc.) are therefore obtained by cumulating $88.32 + 0.05 + 0.1 + 0.1 + 0.1 \ldots$ until 12 monthly items, July to the next June, have been obtained. The June, 1927, item plus 0.05 should, of course,
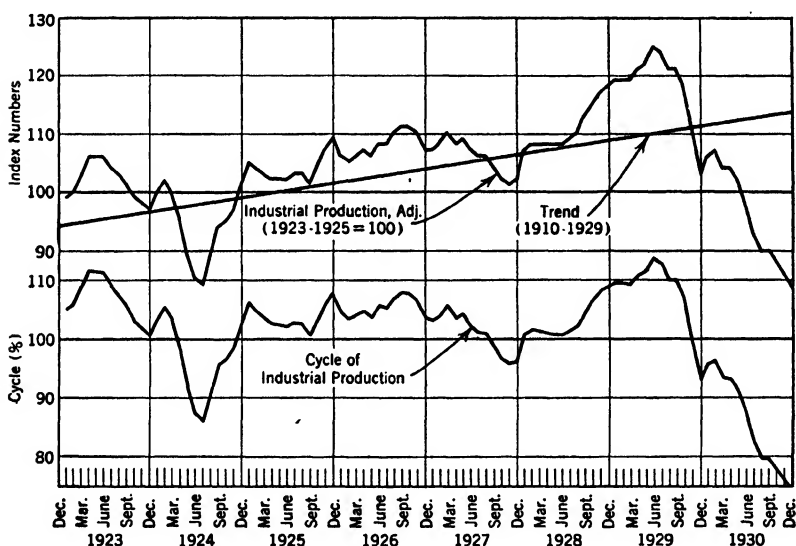
Fig. 13·3.—Monthly Index Numbers of Industrial Production in the United States, 1923–1930, Adjusted for Seasonality, together with Straight-Line Trend Projected from the Base Period, 1910–1919, and Cycle in Percentage Units (see Tables 13·1–13·5.)   (The 1940 revision of this series is not used here because it extends back only to 1923).

check with the 1927 annual trend point. In the same way, monthly trend points for other years in the same parabolic trend may be calculated, whether the annual rise is positive or negative.

The adequacy of such an interpolation may be checked on a chart, and if it seems necessary, semi-annual (or quarterly) trend points may be calculated for the parabola fitted to annual

data (x taken at -2, -1.5, -1, etc.) and the required items may be interpolated on a linear basis, as just explained. Or, if greater accuracy is required, the monthly items may similarly be calculated by the parabolic equation itself (x taken at $\frac{1}{24}$, $\frac{3}{24}$, $\frac{5}{24}$, etc., in both positive and negative directions from the origin).

In calculating the trend, it is well to carry several more decimal places than will be finally used, because in successively adding or subtracting one-twelfth of b the error involved in rounding the decimals may accumulate to a significant figure. When the trend items are read from the machine they may be shortened to approximately the number of significant figures that appear in the data.

The composite cycle.—In historical studies of the business cycle, and occasionally in current studies, it is desirable to weld into a composite the cyclical measures of a number of pertinent individual series. If the series were suitably proportioned samples from various fields of production, they might simply be added after adjustment for seasonality. The cycle might then be isolated as already illustrated for industrial production. But often they are heterogeneous elements of varying importance. The degree of cyclical change may then very well be reduced in each series to such an abstract common denominator as an average deviation or standard deviation, and the composite measure may be found as a weighted average of these ratios (d/AD or d/σ). The procedure as applied to a single month is illustrated in Example 13·2.

In explanation of this example, it may be said that the columns $C\%$, d, AD, and d/AD represent steps already explained in connection with Example 13·1 and Tables 13·2– 13·5. It is assumed that the procedure there described has been carried through several years for the five series to be combined. The example itself (13·2) represents merely the process of obtaining a composite figure for a single month (January, 1929). The weights in the next to the last column represent a considered judgment regarding the relative validity and importance of each series. They take into account the market volume, the accuracy of reporting, and the representative nature

of each series. They are used in computing a weighted average of the d/AD items of each month in building the composite index. It should be noted that the result is expressed in aver-

EXAMPLE 13·2

THE COMPOSITE CYCLE, AD UNITS

Data: Estimated market data for January, 1929.

I. Computed in average deviation (AD) units.

Series	$C\%$	d	AD	d/AD	Weights	Product
Automobile production	145.6	+45.6	23.6	+1.93	3	+5.79
Cotton consumption	114.2	+14.2	8.8	+1.61	4	+6.44
Electric-power production	103.8	+ 3.8	3.7	+1.03	3	+3.09
Freight-car loadings	102.7	+ 2.7	4.5	+0.60	6	+3.60
Pig-iron production	109.2	+ 9.2	18.8	+0.49	4	+1.96

$$20 \,)\,\overline{+20.88}$$

Composite: +1.044 (AD)

II. Computed in percentage $(C\%)$ units.

Series	$C\%$	Weights	Product
Automobile production	145.6	3/23.6 = 0.12712	18.50867
Cotton consumption	114.2	4/8.8 = 0.45455	51.90961
Electric-power production	103.8	3/3.7 = 0.81081	84.16208
Freight-car loadings	102.7	6/4.5 = 1.33333	136.93299
Pig-iron production	109.2	4/18.8 = 0.21277	23.23448

$$2.93858\,)\,\overline{314.74783}$$

Composite: 107.1 (%)

age deviation units. Thus, in this case, the resulting average (+1.044 or 107.1 per cent) implies that, in the given month, business was above normal by 1.044 times the average deviation of the cyclical change measured, or by 7.1 per cent.[1] This con-

---

[1] Since the AD's are used inversely in calculating the composite cycle (Example 13·2), they enter as a harmonic mean into an item expressing the cycle in AD units (+1.044). The harmonic mean of the five AD's, as weighted, is

$$HM = 20 \div \left(\frac{3}{23.6} + \frac{4}{8.8} + \frac{3}{3.7} + \frac{6}{4.5} + \frac{4}{18.8} \right) = 6.806$$

The composite percentage deviation of the cycle from normal is, therefore,

Composite $d\% = + 1.044 \times 6.806 = 7.11$

which necessarily agrees with the $C\%$ item (107.11) obtained in Example 13·2

clusion is much the same as that reached on the basis of industrial production alone (+1.06).

In practice, of course, this manipulation must frequently be applied to a long series of months, and various short cuts are available. Thus, since the weight attached to each component series and the AD for that series are constants, time may be saved by first dividing the weight by the AD and then multiplying each of the deviations throughout that series by the quotient thus obtained. The same procedure may be applied to each of the series. The composite index for each month is merely the sum of these products divided by the sum of the weights.

Conclusion.—Finally, it may be pointed out that the practical purpose of the study of cyclical variations, and of business cycles in particular, is chiefly that of forecasting them or perhaps ultimately of controlling them. It is obvious that if they are to be dealt with their characteristics must be accurately known. The statistician's task, therefore, is primarily that of measuring cyclical changes and their interrelationships. This type of measurement will be considered in Chapter XVIII.

READINGS

(See also special and general references, pages 591 and 597.)

American Statistical Association's Proceedings, Supplement 165A, March, 1929, pp. 152–173. (Papers by Olin W. Blackett, John H. Cover, George R. Davies, H. Mitchell, F. E. Richter, and Donald S. Thompson.)

ANDERSON, MONTGOMERY D., "Measurement of Business Activity in Florida," *University of Florida Economic Series*, 1 (2), May, 1931.

——"An Agricultural Theory of Business Cycles," *American Economic Review*, 21 (3), September, 1931, pp. 427–449.

BLACKETT, O. W., "Measures of Business Conditions in Michigan," *Michigan Business Studies*, 1 (7), July, 1928, pp. 1–44.

——"Regional Indexes of Trade," *Proceedings of the American Statistical Association*, 24 (165–Supplement), March, 1929, p. 164.

part II. The algebraic identity of the two methods may be shown by the two expressions

$$\text{Composite } d\% = \frac{\Sigma(dw/AD)}{\Sigma w} \times \frac{\Sigma w}{\Sigma(w/AD)} = \frac{\Sigma(100 + d)w/AD}{\Sigma(w/AD)} - 100$$

which may readily be shown to be identical.

CHAPMAN, H. H., "A Method of Determining the Sum Which Can Be Invested Profitably in Reforestation for a Sustained Yield," *Journal of Forestry*, 27 (4), April, 1929, pp. 371–374.

COVER, JOHN H., "The Significance of Regional Business Analysis," *Proceedings of the American Statistical Association*, 24 (165–Supplement), March, 1929, pp. 152–155.

ELLSWORTH, D. W., "Revision of the Indices of Business Activity," *Annalist*, 38 (840), February 22, 1929, pp. 388–396.

KUZNETS, SIMON, "Random Events and Cyclical Oscillations," *Journal of the American Statistical Association*, 24 (167), September, 1929, pp. 258–275.

MAVERICK, L. A., "Cycles in Real Estate Activity," *Journal of Law and Public Utility Economics*, May, 1932, pp. 191–199, and February, 1933, pp. 52–56.

SMITH, JAMES G., and HINRICHS, A. F., "Economic Statistics," *Proceedings of the American Statistical Association*, 24 (165–Supplement), March, 1929, pp. 138–143.

TRUESDELL, LEON E., "Pitfalls in the Field of Statistical Forecasting," *Journal of the American Statistical Association*, 33 (202), June, 1938, pp. 373–379.

WHITE, A. E., "Research and Business Cycles," *Certified Public Accountant*, 12 (3), March, 1932, pp. 165–167.

YNTEMA, THEODORE, O., "The Influence of Dumping on Monopoly Prices," *Journal of Political Economy*, 36 (6), December, 1928, pp. 686–698.

EXERCISES AND PROBLEMS

A. EXERCISES

1. On the assumption that the normal of the annual indexes of business listed in Exercise 10·3, page 244, is appropriately represented by a straight-line trend, compute the percentage cycle, Y/T.

2. On the assumption that straight-line trends are suitable, calculate the normal for Exercise 12·2, page 301, as TS, and find the percentage cycle as Y/ST.

3. On the assumption that straight-line trends are suitable, calculate the normal for Exercise 12·3, page 301, as TS, and find the percentage cycle as Y/ST.

ANSWERS TO EXERCISES

1. (a) 85.5, 109.2, 106.6, 112.6, 85.9.
   (b) 116.7, 86.4, 93.5, 91.7, 112.0.
   (c) 102.2, 97.8, 98.9, 100.0, 101.1.
   (d) 95.5, 93.8, 122.7, 96.4, 90.9.
   (e) 93.6, 103.6, 101.8, 110.9, 90.0.
   (f) 89.8, 97.1, 123.4, 93.8, 95.2.
   (g) 100.0, 98.1, 103.7, 98.1.
   (h) 97.4, 102.6, 97.5, 102.4, 104.8, 95.3.
   (i) 97.8, 103.1, 97.0, 101.9, 102.8, 97.3.
   (j) 101.8, 98.2, 101.8, 98.1, 96.2, 103.8.
   (k) 97.2, 102.9, 102.0, 96.8, 103.3, 97.7.

(*l*) 101.7, 98.3, 101.8, 98.2, 96.4, 103.7.
(*m*) 97.3, 102.8, 101.9, 97.0, 103.1, 97.8.
(*n*) 100.0, 85.7, 121.4, 92.9, 114.3, 78.6, 107.1.
(*o*) 98.3, 100.8, 102.4, 100.0, 99.2, 97.6, 101.6.
(*p*) 101.6, 97.5, 99.2, 100.0, 102.5, 100.8, 98.3.

2. (*a*) $a = 100.0$, $b = 9.5$.   (*e*) $a = 18.0$, $b = 8.0$.
(*b*) $a = 100.0$, $b = 2.0$.   (*f*) $a = 212.5$, $b = 80.0$.
(*c*) $a = 100.5$, $b = 4.0$.   (*g*) $a = 60.0$, $b = 8.0$.
(*d*) $a = 100.0$, $b = 8.0$.

Percentage cycles:

(*a*) 102.2, 101.2, 99.2, 97.1, 96.7, 97.6, 100.3, 101.5, 101.7, 101.8, 100.4, 96.4.

(*b*) 102.8, 98.1, 100.3, 98.9, 98.6, 101.4, 99.3, 100.5, 100.9, 98.4, 101.1, 99.5.

(*c*) 102.6, 101.3, 97.6, 99.0, 97.3, 100.1, 101.8, 99.6, 101.1, 99.0, 98.8, 100.2.

(*d*) 100.0, 100.2, 99.9, 99.9, 99.9, 100.0, 100.1, 100.0, 99.9, 99.9, 100.3, 100.1.

(*e*) 91.7, 87.3, 102.8, 103.7, 99.8, 99.1, 99.2, 98.4, 102.3, 103.2, 97.7, 96.0.

(*f*) 96.9, 98.6, 99.1, 97.7, 98.0, 99.4, 99.7, 98.2, 98.4, 99.7, 100.0, 98.4.

(*g*) 99.7, 99.4, 100.4, 100.2, 99.9, 99.9, 99.9, 100.0, 100.2, 100.3, 99.5, 99.8.

3. (*a*) $a = 100.0$, $b = 9.5$.   (*d*) $a = 199.0$, $b = -8.0$.
(*b*) $a = 100.0$, $b = 8.0$.   (*e*) $a = 201.0$, $b = 8.0$.
(*c*) $a = 200.0$, $b = -8.0$.

B. PROBLEMS

4. Utilizing the data of Problem 5, page 302, compute a straight-line trend, adjusted to months, and find the percentage cycle.

5. Utilizing the data of Problem 6, page 303, compute a straight-line trend, adjusted to months, and find the percentage cycle.

6. Consult the *Survey of Current Business* for further data, and prepare indexes of business cycles.

7. The following figures represent approximate percentage deviations from normal of the three series indicated, the time unit being 6 months. Reduce each series to units of its own average deviation, and plot.

Year	Stocks (1)	Business activity (2)	Interest (3)
1903	2	6	2
	− 8	2	7
1904	−14	− 8	6
	− 9	−14	−10
1905	− 5	− 4	−14
	9	− 2	− 8
1906	16	5	3
	12	8	7
1907	10	15	11
	− 5	13	12
1908	−21	− 8	11
	− 6	−17	− 9
1909	4	− 9	−18
	10	− 2	−13
1910	9	11	3
	− 4	4	10

8. Reduce to average deviation and standard deviation units the cycles indicated in the following data representing wholesale prices and interest rates, and plot the results.

WHOLESALE PRICES

Quarter	1909	1910	1911	1912	1913
1	−12	11	− 5	− 1	7
2	−10	5	−11	6	−1
3	− 4	2	− 5	7	−3
4	8	− 1	− 2	11	−2

INTEREST RATES

Quarter	1909	1910	1911	1912	1913
1	− 3	13	− 6	−15	10
2	− 2	14	−19	−10	17
3	− 6	18	−15	− 2	10
4	12	3	−21	5	− 3

PART II

CHAPTER XIV

SIMPLE CORRELATION

The term "correlation" is defined in popular usage as meaning "similarity" or covariation. Thus, it may be said, for instance, that there is "close correlation" between price levels and business activity, or that there is little correlation between changes in costs and in prices. In statistics, the term is similarly used to refer to the concurrent variation of two or more series or variables. However, its statistical usage is manifold, for the same term "correlation" is used to refer to the actual covariation, to the technique or procedure of measuring its extent and direction, and to the description of that measurement. Consideration must be given to each of these usages.

Some simple illustrations may suffice to clarify the nature of the variety of relationships described as correlation. It is widely recognized, for instance, that corn yields are closely related to rainfall. If rainfall is light, yields are reduced. As rainfall increases (up to the point where it affects crops adversely) the yield is also increased. Covariation is clear and within these limits is positive, i.e., as one variable increases, the other gains also. Again, in many periods, as costs of living have risen, real wages have fallen. Here there is covariation, but it is *negative*, i.e., the one variable rises as the other falls. Numerous similar illustrations may be found in a wide range of business and economic activity.

A number of methods of analysis have been devised to measure the extent and nature (whether positive or negative) of correlation. In effect, each of these methods seeks to discover how much of the variation in one variable can be accounted

for or explained by the variability of the correlated variable. A major portion of this and several succeeding chapters is devoted to description of these techniques and procedures.

Uses of correlation.—It will be apparent that the process or technique thus described has many possible uses in the statistical analysis of business conditions, for it is frequently worth while to note and measure covariation in business data. It is significant to note, for instance, how prices of certain commodities vary with changes in visible supplies, or with fluctuations in demand, or with variations in production of such goods in immediately preceding periods. Again, management may seek to discover how sales or production vary with differences in certain characteristics of salesmen or workers involved, such as education, special training, age, physical and mental characteristics, and numerous other such features.

In other cases, the techniques of correlation may be used to measure covariation in bond yields with such features of these securities as their prices, their terms, and more complicated characteristics; or stock prices and distinctive characteristics of the corporations they represent; or insurance losses and peculiar features of the risks involved. In all these and numerous similar situations, the methodology of correlation analysis is appropriate.

Limitations of correlation.—Because considerably more is frequently read into the results of correlation analysis than the process actually implies, it may be well at the outset to indicate some important limitations. In the first place, no measure of correlation, however great, proves the existence of a causal relationship between the variables that are compared. Their similarity may be a matter of chance or of remote relationships to other changing series. The interpretation of any measure of correlation, however imposing, must therefore be supplied, not from the mere statistical analysis, but from an understanding of common causes or other relationships between the phenomena correlated.

It is also important to recognize the fact that an association that is real and measurable within the limits of the observations used in measuring it may not prevail outside those limits. Thus

a measurable similarity in the way corn yields and rainfall vary within the limits of normal rainfall and normal yields will not characterize extremes in which there is an almost entire absence of rainfall, or a series of floods. For this reason, conservative statistical procedure requires that a measure of correlation should be regarded as valid only within the range of observations.

Linear or rectilinear correlation.—It has been said that the correlation technique may be applied to two or more variables. When more than two variables are included, the process is said to involve *partial* or *multiple* correlation, or both, and the procedure becomes sufficiently complex to deserve special attention and treatment. Moreover, certain types of correlation procedure that may be described as *non-linear* or *curvilinear* are also distinctive enough to require special consideration. Multiple and partial correlation are considered in a later chapter, and another chapter is devoted to non-linear relationships and to several special types of correlation. In the present chapter, unless reference is made specifically to these more complicated procedures, the term *correlation* is used to refer to the simplest measurement of relationship between two variables, the so-called *linear* (or, more strictly, *rectilinear*) correlation.

Correlation analysis illustrated.—The method of correlation analysis may be explained by reference to a simple illustrative problem, such as that suggested by Table 14·1. The table summarizes data with respect to weekly sales (expressed in thousands of dollars and assumed to represent the efficiency of the respective salesmen) compared with scores made by these salesmen on a test administered by the concern at the time they were employed. The data are, of course, hypothetical and have been made extremely simple and limited in number in order to facilitate explanation of correlation theory and procedure.

The immediate purposes or objectives of such correlation analysis are several: (1) it seeks to secure a quantitative appraisal of the covariation, a measure of the extent to which the test appears to select those who are successful salesmen; (2) it provides a basis upon which this measured covariation may be appraised as to its significance, i.e., it permits a comparison with accidental or chance covariation; (3) it facilitates

prediction as to the probable degree of success to be achieved by new candidates for sales positions. In other words, correlation analysis first results in a measure of covariation. This measure may be compared with those that might appear by chance or accident, to discover whether the covariation is really significant. It then provides a predicting equation, by means of which it is possible to forecast sales for various test scores.

TABLE 14·1

COMPARISON OF ADMISSION-TEST SCORES AND SALES

Salesmen	Admission-test scores	Weekly sales (in thousands of dollars)
A	4	5
B	5	4
C	6	5
D	4	6
E	5	9
F	6	10
G	6	9
H	7	12
I	9	11
J	8	9

The essential nature of these objectives and the problems they involve may be seen more clearly from a chart such as Fig. 14·1, where sales (designated as Y) are plotted against test scores (designated as X). In correlation procedure, it is customary to plot what is regarded as the independent variable on the X scale, with values increasing from left to right, and the dependent variable on the Y scale, with values increasing from bottom to top. It will be seen at once that there is a considerable degree of covariation in the two series, a fact clearly indicated by the manner in which the Y values increase as X values grow. If, however, there were perfect linear correlation or covariation, the points representing the individuals would fall in a straight line, for each increase in test scores would be

accompanied by a proportionate increase in sales. The fact that actual data do not represent such a line indicates that linear covariation is not perfect, that there is not a uniform relationship throughout both series.

Understanding of the nature of linear correlation may be furthered by consideration of Fig. 14·2, in which several possible patterns are compared.
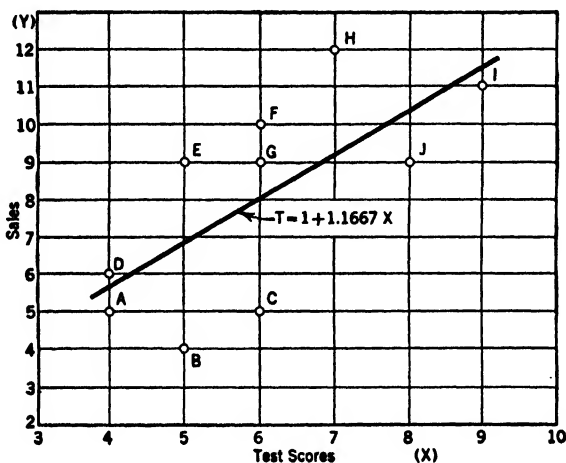

FIG. 14·1.—The Linear Regression of Sales (Y) on Test Scores (X). (Data: See Table 14·1)

The fact that, in the data described in preceding paragraphs, the pattern does not represent a straight line running from the lower left hand to the upper right hand in effect states the first ·problem of correlation, which is the measurement of the extent to which covariation does prevail. That question may be restated: to what extent do the series vary together, or how much correlation features their association?

Pearson product-moment r.—The basic practical question to be answered may be restated as follows: when the test scores of individuals are above or below the mean of such scores, are the sales of the same individuals correspondingly above or below the mean of sales? That is, does the man making low test scores also measure low in salesmanship, and does the candidate ranking high in test scores also rank high in sales-

manship? Finally, the question continues: to what extent does this relationship prevail?

The answer to this question is found in a measure known as the Pearsonian *coefficient of correlation* (symbol r), which may be described by the formula

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$
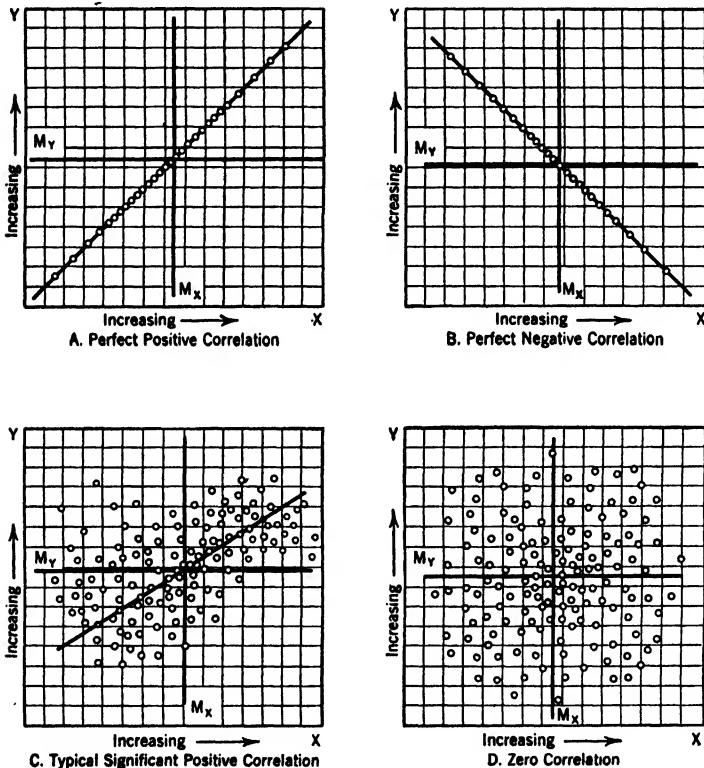

Fig. 14·2.—Patterns of Linear Correlation. (Data assumed for purposes of illustration.)

in which Σxy is the usual summation of products of X and Y deviations from the means of the respective series ($x = X - M_X$ and $y = Y - M_Y$) already described in connection with the fitting of least-square linear trends, and N is the number of cases involved. The standard deviations are calculated as

usual. It will readily be seen that $N\sigma_x\sigma_y = \sqrt{\Sigma x^2 \Sigma y^2}$, since

$$N\sigma_x\sigma_y = N\sqrt{\frac{\Sigma x^2}{N}}\sqrt{\frac{\Sigma y^2}{N}} = N\sqrt{\Sigma x^2 \Sigma y^2}\sqrt{\frac{1}{N^2}} = \sqrt{\Sigma x^2 \Sigma y^2}$$

where N and $\sqrt{1/N^2}$ cancel each other. The second form is usually more convenient in laboratory practice.

This formula, as will be explained more fully in another connection, represents a ratio of one variability to another, the two variabilities being (1) the variability of the trend items in a straight-line trend fitted by the method of least squares to the Y items, and (2) the variability of the Y data themselves to which this trend is fitted, each such range of variation being measured in standard deviation units. The ratio is thus [1]

$$r = \frac{\sigma_t}{\sigma_y} ; \quad \text{or} \quad r^2 = \frac{\sigma_t^2}{\sigma_y^2} = \frac{\Sigma t^2}{\Sigma y^2}$$

where σ_t is the standard deviation of the trend items ($\sigma_t = \sqrt{\Sigma t^2 \div N}$, where $t = T - M_Y$) and σ_y is similarly calculated from the Y items. (Note that M_T and M_Y are identical.)

The formula, $r^2 = \sigma_t^2/\sigma_y^2$, defines the squared coefficient of correlation as the ratio of regression variance, σ_t^2, to data variance σ_y^2 (or, more conveniently, $\Sigma t^2/\Sigma y^2$), and is theoretically the most important correlation formula. It is, in fact, the general formula of correlation, as will appear later. If used in simple correlation, it is generally written as its algebraic equivalent

$$r^2 = \frac{b\Sigma xy}{\Sigma y^2}$$

where b is the slope of the trend or regression line fitted to the Y data.

---

[1] That is, $r^2 = \sigma_t^2/\sigma_y^2$, which (multiplying each term by N) equals $\Sigma t^2/\Sigma y^2$. The identity of this equation with the original equation for r is shown as follows: The trend equation, $T = a + bX$, when centered as deviations from M_Y, becomes $t = bx$. Squaring and summing this equality,

$$\Sigma t^2 = b^2\Sigma x^2 = \left(\frac{\Sigma xy}{\Sigma x^2}\right)\left(\frac{\Sigma xy}{\Sigma x^2}\right)\Sigma x^2 = \frac{(\Sigma xy)^2}{\Sigma x^2}$$

Hence, the ratio $\Sigma t^2/\Sigma y^2 = (\Sigma xy)^2/\Sigma x^2\Sigma y^2$, which is r^2 as previously expressed. Note that Σt^2 may also be written $b\Sigma xy$.

<div align="center">

EXAMPLE 14·1

MEASUREMENT OF CORRELATION [1]

</div>

Data: Psychological test scores and average weekly sales of a group of salesmen.

Sales-men	Test scores X	Sales (in thousands of dollars) Y	Devia-tions $(X-M_X)$ x	Devia-tions $(Y-M_Y)$ y	x^2	xy	y^2
A	4	5	-2	-3	4	6	9
B	5	4	-1	-4	1	4	16
C	6	5	0	-3	0	0	9
D	4	6	-2	-2	4	4	4
E	5	9	-1	1	1	-1	1
F	6	10	0	2	0	0	4
G	6	9	0	1	0	0	1
H	7	12	1	4	1	4	16
I	9	11	3	3	9	9	9
J	8	9	2	1	4	2	1
	60	80			24	28	70

Computation of r:

$$M_X = \frac{\Sigma X}{N} = \frac{60}{10} = 6$$

$$M_Y = \frac{\Sigma Y}{N} = \frac{80}{10} = 8$$

$$\sigma_x^2 = \frac{\Sigma x^2}{N} = \frac{24}{10} = 2.4; \sigma_x = \sqrt{2.4} = 1.5492$$

$$\sigma_y^2 = \frac{\Sigma y^2}{N} = \frac{70}{10} = 7.0; \sigma_y = \sqrt{7.0} = 2.6458$$

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{28}{10 \times 1.5492 \times 2.6458} = 0.683$$

or

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{28}{\sqrt{24 \times 70}} = \frac{28}{40.9878} = 0.683$$

---

[1] As is explained more fully in the next chapter, r may be found by the use of centering formulas, such as

$$\Sigma x^2 = \Sigma X^2 - M\Sigma X$$
$$\Sigma xy = \Sigma XY - M_X\Sigma Y \quad \text{or} \quad \Sigma XY - M_Y\Sigma X$$

In simple correlation, however, the formula first mentioned, particularly in the second form $(r = \Sigma xy/\sqrt{\Sigma x^2 \Sigma y^2})$, is most commonly used. The elementary process in which this formula is applied is illustrated in Example 14·1. There the X values are summarized in one column, the paired Y values occupying another. The mean of each series is found, and then this mean is subtracted from each item in its series to provide values of x and y. These deviations are than squared, as shown, and the cross products, xy, are found. Totals are substituted in the formula as illustrated, and the measure of correlation is found to be $r = 0.683$. Though the method of finding r thus described is not the most convenient, it indicates clearly the nature of simple correlation.[1]

Assumptions of correlation procedure.—The assumptions upon which this measure of covariation is based and which explain its meaning are of utmost significance in understanding the nature of the coefficient of correlation thus obtained. The measurement of covariation by the coefficient actually involves two principal steps which, however, are obscured by the manipu-

---

as illustrated with the following X and Y data:

X	Y	X^2	XY	Y^2
2	1	4	2	1
4	7	16	28	49
4	9	16	36	81
10	15	100	150	225
—	—	—	—	—
20	32	136	216	356

$M_X = 5$ $M_Y = 8$ $M_X \Sigma X = 100$ $M_X \Sigma Y = 160$ $M_Y \Sigma Y = 256$

$\Sigma x^2 = 36$ $\Sigma xy = 56$ $\Sigma y^2 = 100$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{56}{\sqrt{36 \times 100}} = \frac{56}{60} = 0.933$$

[1] It may be noted that xy is, in theory, an abstraction. The Y scale is, in this conception, measured in y/σ_y units and the X scale in x/σ_x units. In these units, r is merely the slope of the regression line; that is

$$r = b\frac{\sigma_x}{\sigma_y} \quad \text{and} \quad b = r\frac{\sigma_y}{\sigma_x}$$

ulation of the data in the formula.   With reference to Fig. 14·1, these steps may be briefly described as follows:

1. The coefficient assumes that the best representation of covariation between the two variables is a least-squares straight line fitted to the data.   In effect, this line is fitted by the same formula and in a manner precisely similar to that utilized in fitting straight-line trends to time series.

2. The coefficient assumes that the best *measure* of covariation is the extent to which such a straight-line " trend " actually represents the data, i.e., the extent to which the data conform to the line thus constructed.   If correlation is high, the data tend to cluster close to this line; if they form no such pattern but are scattered generally throughout the ranges of the two series, then the line is a poor representation of the data, and correlation is small.   The variation in the fitted " trend " line as measured by σ_t^2 is sometimes described as " accounted-for " variance.   Hence, since the squared coefficient is the ratio of variance in the straight line itself (T) to the variance of the sales data (Y) as measured by the respective squared standard deviations, it is the ratio of " accounted-for " variance to total variance.   It is entirely possible to calculate the coefficient by following these steps through, as will be seen in later paragraphs.

Significance of the coefficient of correlation.—A full explanation of the basis upon which the significance of such a measure of correlation is appraised must await further explanation of the correlation process.   However, it is possible to describe at this point the most satisfactory method of evaluating that significance.   Tables and charts are available showing the highest coefficient of correlation that might appear by chance once in 20 times (5 per cent level) and once in 100 times (1 per cent level) in N pairs of samples drawn from normally distributed X and Y universes.   If the coefficient found is greater than that which might appear once in 20 times by chance but not so great as the 1 per cent value, it is said to be *significant*. The charts to be used for this purpose appear on pages 559–560. It will be noted that, for 10 pairs of items, the least *significant* value is approximately 0.63, while the least highly significant

value is approximately 0.76. The coefficient found in this case, $r = 0.683$, is, therefore, *significant* but not *highly significant*.[1]

Such tests of significance are based on what is known as the null hypothesis, so called because it assumes conditions directly opposed to those sought in the analysis, and it is the purpose of such analysis to disprove the hypothesis. Thus, as a test of the significance of correlation measures, it is assumed that no correlation exists, and chance values of r or related measures are calculated on this assumption. These values are then used as criteria in appraising actual discovered values. Thus 5 per cent and 1 per cent probable values based on the null hypothesis are used as limits in evaluating actual coefficients.

<div align="center">

READINGS
</div>

<div align="center">

EXERCISES AND PROBLEMS

A. EXERCISES
</div>

1. Making use of the data summarized below, calculate the coefficients of correlation r_{01}, r_{02}, and r_{12}. Are the measures statistically significant? Note that r_{01} implies the correlation of X_0 as dependent, with X_1 as independent.

CASE	X_1	X_2	X_0
A	10	9	21
B	6	4	10
C	9	6	11
D	10	9	13
E	12	11	21
F	13	13	22
G	11	8	12
H	9	4	10

[1] A more general method of evaluation applicable to r is available in the Table of F (cf. page 586). To use this table it is necessary to calculate

$$F = \frac{r^2}{1 - r^2} \times (N - 2)$$

which in the case of Example 14·1, where r was found to be 0.683, is

$$F = \frac{0.683^2}{1 - 0.683^2} \times (10 - 2) = \frac{0.466}{0.534} \times 8 = 6.98$$

In the first column of the table of F, in row $N - 2 = 8$, the 5 and 1 per cent levels of the sampling distribution of F are given as 5.32 and 11.26, respectively. Hence, as before, r is evaluated as significant, but not highly significant. A discussion of the statistic F appears in the Appendix, pages 554–556.

2. Calculate the coefficients of correlation r_{01}, r_{02}, and r_{03} for the data listed below:

CASE	X_1	X_2	X_3	X_0
A	14	20	18	13
B	6	6	7	6
C	10	11	8	8
D	12	28	10	11
E	14	31	18	13
F	20	32	19	14
G	12	25	9	13
H	8	7	7	2

3. The following data are assumed to represent three independent series (X_1, X_2, and X_3) and a dependent series (X_0).

Compute the correlations, r_{01}, r_{02}, r_{03}, r_{12}, r_{13}, and r_{23}.

CASE	X_1	X_2	X_3	X_0
A	10	25	11	26
B	6	11	16	15
C	9	16	14	16
D	10	33	11	18
E	12	36	9	26
F	13	37	7	27
G	11	30	12	17
H	9	12	16	15

4. From the following data calculate r_{01}, r_{02}, and r_{12}.

CASE	X_1	X_2	X_0
A	31	25	42
B	20	21	34
C	21	24	38
D	23	25	40
E	31	27	42
F	32	28	48
G	22	26	40
H	20	24	36

5. From the following data calculate r_{01}, r_{02}, r_{03}, and r_{04}.

CASE	X_1	X_2	X_3	X_4	X_0
A	12	8	20	16	18
B	15	16	8	19	21
C	6	12	19	15	16
D	16	24	13	18	29
E	11	10	15	13	19
F	7	2	18	14	10
G	17	26	12	17	27

ANSWERS TO EXERCISES

1. $r_{01} = 0.7375$; $r_{02} = 0.8750$; $r_{12} = 0.8958$. Least significant $r = 0.707$; least highly significant $r = 0.834$.

2. $r_{01} = 0.8125$; $r_{02} = 0.8812$; $r_{03} = 0.7562$.

3. $r_{01} = 0.7375$; $r_{02} = 0.7250$; $r_{03} = -0.8750$.
  $r_{12} = 0.8812$; $r_{13} = -0.8958$; $r_{23} = -0.9458$.

4. $r_{01} = 0.8625$; $r_{02} = 0.9062$; $r_{12} = 0.7375$.

5. $r_{01} = 0.8869$; $r_{02} = 0.9524$; $r_{03} = -0.6071$; $r_{04} = 0.6548$.

B. PROBLEMS

6. The following summary compares labor turnover rates (reduced to unavoidable separation rates) with average daily earnings for a number of groups of workers. The management seeks to discover, by this type of analysis, what employees are responsible for excessive proportions of turnover (data reduced for classroom purposes). The groups are of equal size.

Group	Average daily earnings	Unavoidable separation rates
A	$ 2.00	1.8
B	11.00	2.0
C	10.00	3.0
D	3.00	1.5
E	9.00	2.5
F	4.00	2.0
G	7.00	2.6
H	5.00	2.1
I	8.00	2.2
J	6.00	2.3

Find: (*a*) the average (*M*) of daily earnings; (*b*) the median separation rate; (*c*) the Pearsonian correlation coefficient; (*d*) the *r* required for 5 per cent significance (groups considered as units).

7. Management seeks to discover a measure of correlation between length of service on the part of a certain type of machine and the annual repair bills on such machines. The following data (shortened for purposes of this problem) are considered:

Machine	Years of service X	Annual repair costs Y
A	1	$2.00
B	3	1.50
C	4	2.50
D	2	2.00
E	5	3.00
F	8	4.00
G	9	4.00
H	10	5.00
I	13	8.00
J	15	8.00

(*a*) Chart the data, designating years of service as the *X* series and annual repair costs as the *Y* series.

(*b*) Find the coefficient of correlation, *r*.

(*c*) Is the measure of correlation significant?

8. In attempting to improve its selection program and thus avoid unnecessary expenditures for training, management seeks a measure of the correlation between training-program scores and high-school grades. The data summarized below represent a sample of those analzyed for this purpose:

Trainee	Average of high-school grades (X)	Average train- ing-course grade (Y)
1	10	9
2	9	9
3	9	10
4	8	7
5	8	8
6	8	7
7	7	5
8	7	5
9	7	4
10	7	6

(a) Secure the Pearsonian coefficient of correlation, r.

(b) Is the coefficient statistically significant?

(c) Plot the data.

9. On the basis of the following tabulations comparing years of service with ratings, management seeks to discover whether or not there is a distinct tendency to rate old employees higher than more recent additions to the working force.

Employee	Service (in years)	Rating	Employee	Service (in years)	Rating
A	1	5	K	6	9
B	9	6	L	7	4
C	8	8	M	1	2
D	3	8	N	1	3
E	3	6	O	3	8
F	2	7	P	1	6
G	4	5	Q	2	5
H	5	6	R	2	3
I	5	4	S	4	4
J	6	5	T	2	7

(a) Chart the data.

(b) Calculate the Pearsonian r.

CHAPTER XV

SIMPLE CORRELATION (Continued)

In the preceding chapter, the essential nature of correlation has been explained, and attention has been directed to some of the simpler methods of measuring covariation between two series. A number of extensions and refinements of the processes described in that chapter may now be given attention. One of the most frequent uses of correlation, for instance, is that in which the known relationship between two series is made the basis for predicting variation in one of them from given values in the other. The method by which this objective is accomplished is described in the early pages of this chapter.

In practice, as has been noted, it is seldom convenient to note deviations of individual items from their mean, for which reason the "crude-data" method described in this chapter will frequently be found useful. In other situations, where grouped data are to be analyzed, modifications of the techniques already described are also desirable. Hence attention is given to special methods of correlation for grouped data. Finally, this chapter considers several specialized types of correlation, adaptable to frequently encountered problems, of which the most important are rank correlation and fourfold correlation.

Prediction and estimation.—The problem of estimation or prediction is solved by fitting a least-squares trend line, or *regression line*, as such a trend is usually called, to the data of Example 14·1, p. 331. The calculation of the trend appears in Example 15·1. For convenience in further analysis, the data are arrayed according to test scores. The customary normal

equations (see page 236) may be used, or the trend may be fitted by formulas derived from these normal equations, thus:

$$T = a + bx$$

$$a_{yx} = \frac{\Sigma Y}{N}; \quad b_{yx} = \frac{\Sigma x Y}{\Sigma x^2} = \frac{\Sigma xy}{\Sigma x^2}; \quad \left(\text{or } b_{yx} = r \frac{\sigma_y}{\sigma_x} \right)$$

EXAMPLE 15·1

LINEAR REGRESSION

Data: Test scores and weekly sales, arrayed according to test scores (see Example 14·1, p. 331).

Sales-men	Test scores	Sales (in thousands of dollars)	$X - M_X$	$Y - M_Y$			
	X	Y	x	y	x^2	xy	y^2
A	4	5	−2	−3	4	6	9
D	4	6	−2	−2	4	4	4
B	5	4	−1	−4	1	4	16
E	5	9	−1	1	1	−1	1
C	6	5	0	−3	0	0	9
G	6	9	0	1	0	0	1
F	6	10	0	2	0	0	4
H	7	12	1	4	1	4	16
J	8	9	2	1	4	2	1
I	9	11	3	3	9	9	9
	60	80	0	0	24	28	70
	$M_X = 6$	$M_Y = 8$					

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{28}{24} = 1.1667, \quad \text{or} \quad b = r\frac{\sigma_y}{\sigma_x} = (0.683)\frac{2.646}{1.549} = 1.1667$$

Origin at $x = 0$ (i.e., at M_X):

$$a = M_Y = 8$$

$$T \text{ or } Y' = a + bx = 8 + 1.1667x$$

Origin at $X = 0$ (crude data):

$$a = M_Y - bM_X = 8 - (1.1667 \times 6) = 1$$

$$T \text{ or } Y' = a + bX = 1 + 1.1667X$$

The equation for the straight line becomes

$$T = 8 + 1.1667x \text{ (origin at } M_X)$$

However, it is frequently more convenient to express the equation in terms of X rather than x, in which case a value of a when $X = 0$ must be found. When X is to be used,[1]

$$a = M_Y - bM_X \text{ (origin at } X = 0)$$

and the equation is

$$T = 1 + 1.1667 X \text{ (origin at } X = 0)$$

where T (often written Y' or Y_c) is the point where the regression line crosses any given X ordinate.

From this equation, trend items corresponding to each individual test score may be readily secured by substituting individual test scores for X. Thus, in further tests, for a test score of 4, the trend or predicted sales value is

$$T = 1 + (1.1667 \times 4) = 1 + 4.67 = 5.67$$

Similarly, the estimated trend for a test score of 6 is 8.00, that for a test score of 9 is 11.50. These are the estimated or predictable values of Y (sales), as indicated in Example 15·2, column 4, that is, the sales that might be expected of a new applicant making the given test score.

If T is plotted against X (see Fig. 14·1, page 328) the trend conforms to a straight line with the height of $a = 1$ at $X = 0$, and with a slope of $b = 1.1667$, as indicated in the equation. The series of points comprising the trend line indicates the estimated efficiency of applicants making various test scores, as it would be predicted on the basis of the regression equation, within the limits here studied. As thus estimated, sales tend to increase by 1.1667 (thousands of dollars) for each point increase in test scores.

---

[1] This formula for a is an adaptation of the first normal equation (see page 236):

$$Na + b\Sigma X = \Sigma Y, \text{ or } Na = \Sigma Y - b\Sigma X$$

hence,

$$a = (\Sigma Y - b\Sigma X) \div N = M_Y - bM_X$$

It will be seen that, if Fig. 14·1 is drawn on a large scale, the probable efficiency of future candidates for jobs may be read from the chart itself.   A given test score is located as a

EXAMPLE 15·2

ESTIMATES BASED ON LINEAR REGRESSION

Data: Test scores and weekly sales (see Example 15·1).

Sales-men	Test scores X	Sales (in thousands of dollars) Y	Estimated sales 1 + 1.167X T	Errors of estimate $d = Y - T$	d^2	$T - M_y$ t	t^2
A	4	5	5.67	−0.67	0.45	−2.33	5.43
D	4	6	5.67	0.33	0.11	−2.33	5.43
B	5	4	6.83	−2.83	8.01	−1.17	1.37
E	5	9	6.83	2.17	4.71	−1.17	1.37
C	6	5	8.00	−3.00	9.00	0	0
G	6	9	8.00	1.00	1.00	0	0
F	6	10	8.00	2.00	4.00	0	0
H	7	12	9.17	2.83	8.01	1.17	1.37
J	8	9	10.33	−1.33	1.77	2.33	5.43
I	9	11	11.50	−0.50	0.25	3.50	12.25
	60	80	80.00 $M = 8$	0.0	37.31	0	32.65

$$\sigma_y^2 = 7; \quad \sigma_y = 2.646 \text{ (see Example 14·1)}$$

$$\sigma_d^2 = \frac{\Sigma d^2}{N} = \frac{37.31}{10} = 3.731; \quad \sigma_d = \sqrt{3.731} = 1.93$$

$$\sigma_t^2 = \frac{\Sigma t^2}{N} = \frac{32.65}{10} = 3.265; \quad \sigma_t \doteq \sqrt{3.265} = 1.81$$

Check:

$$\Sigma t^2 = b\Sigma xy = 1.1667 \times 28 = 32.67$$

$$\Sigma d^2 = \Sigma y^2 - \Sigma t^2 = 70.00 - 32.67 = 37.33[1]$$

point on the X scale.   From this point, by reading directly up to the trend line and left to the Y scale, the Y value that repre-

[1] For proof of this relationship of Σd^2 and Σt^2, see Appendix, page 535.

sents the estimated sales is found. Or this calculation may be made directly from the trend equation, without reference to the graphic representation, as has been illustrated.

Standard error of estimate.—It will be noted, in Example 15·2, that, although variation in test scores accounts for some of the variation in sales, it does not account for all of it, as is clearly shown in the column described as " errors of estimate " (symbol d). These errors represent the difference between actual sales (Y) and those estimated (T) for each individual on the basis of the regression equation, i.e., $d = Y - T$, and Σd is necessarily zero. If the test were perfect as a means of estimating sales ability, there would be no errors of estimate, and trend values would be identical with actual values. To the extent that the test is imperfect, the variability of the actual data is greater than that of the estimated or trend values, and errors of estimate appear. The standard deviation of these residuals or errors of estimate is the most useful measure of the failure of the regression formula to provide an exact basis for estimating values of the dependent series from given values in the independent series. This measure is called the *standard error of estimate* (σ_d or simply S). As has been indicated in Example 15·2, it may be calculated directly as the standard deviation of the individual errors of estimate,[1]

$$\sigma_d^2 \text{ or } S^2 = \frac{\Sigma d^2}{N} = \frac{37.3}{10} = 3.73$$

$$\sigma_d = \sqrt{3.73} = 1.93$$

In practice, however, the calculation of individual estimated t values and individual errors of estimate is unnecessarily cumbersome, and the standard error of estimate is generally found by formula as:

$$\sigma_d \text{ or } S = \sigma_y\sqrt{1 - r^2} = 2.65\sqrt{0.533} = 1.93$$

[1] This calculation is for the sample only. The best estimate of S for the universe would make use of $N - 2$ degrees of freedom, thus

$$\sigma_d = \sqrt{\Sigma d^2 \div (N - 2)}$$

which is the standard error of estimate corrected for sampling.

Since the "accounted-for" and the "unaccounted-for" variances together equal the total variance to be explained, it is always true that, for the given sample,

$$\Sigma t^2 + \Sigma d^2 = \Sigma y^2, \quad \text{and} \quad \sigma_t^2 + \sigma_d^2 = \sigma_y^2$$

The standard error of estimate is useful in evaluating the inaccuracy of prediction based on the regression equation. When corrected for sampling it measures the expected variability of estimated values from true values, assuming that data
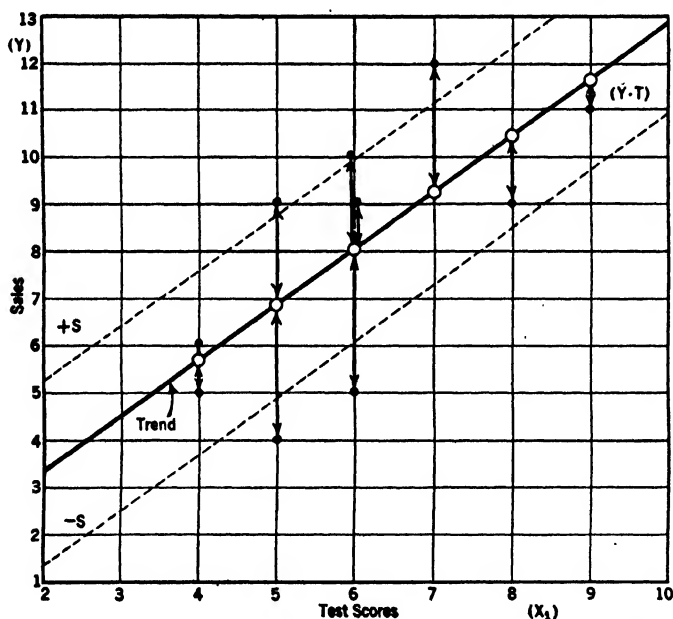

Fig. 15·1.—The Standard Error of Estimate (S or σ_d) of the Linear Regression Shown in Fig. 14·1.

are adequate and represent random samples from normal distributions. Under such conditions of large-sample theory, it may be roughly estimated that all items will fall within a range of three standard errors of the regression line. The limits of one standard error of estimate are shown in Fig. 15·1, where the data are those utilized in earlier calculations. The standard error of estimate is measured on the vertical Y scale.

Prediction reversed.—In certain types of analysis, it may be desirable to reverse prediction, to estimate the probable value of X from a given Y value. For example, in the problem already discussed, it might happen that the sales record of a certain salesman was known, and question might arise concerning the probable score he would make in the psychological entrance test.

At first glance, one might assume that the regression equation used as a basis for estimating Y from given values of X might be equally effective as a basis for forecasting X values from given Y values. This would mean that, if, as previously estimated, a test score of 7 means a probable sales record of 9.17, then a sales record of 9.17 would be the basis of predicting a test score of 7. This appears plausible but is not true. The regression of Y on X has been fitted, according to the criterion of least squares, in such a manner that the d_Y variance, $\Sigma(Y - T_{yx})^2/N$, as measured on the Y scale, is at a minimum. A regression to be used in predicting X values and determining the standard error of estimate in such predictions must provide a least-squares trend based on a minimum d_X variance, $\Sigma(X - T_{xy})^2/N$, as measured on the X scale. Unless, therefore, there is perfect correlation, there are always two regressions: the one a regression of Y on X, used in predicting Y values from given X values; the other the regression of X on Y, used for reverse prediction.

Obviously, reversed prediction could be calculated by simply reversing the X and Y labels applied to the data, that is, in the case at hand, by labeling sales records (X) and test scores (Y), and recomputing the whole problem. This procedure is not necessary, however, since the same result may be attained by leaving the labels as they are and calculating a new regression equation in which Y takes the place of X and new values of a and b are utilized. The amount of additional calculation is not extensive.

It is obvious that the degree of correlation is the same in either case, since interchanging the X's and Y's in the formula for r makes no difference. But the formulas for the regression equation are different, i.e. (the second subscript attached to a

or b indicates the independent element, or base, from which it is calculated):

(1) When X is regarded as independent:

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = r \frac{\sigma_y}{\sigma_x}$$

$$a_{yx} = M_Y - b_{yx}M_X$$

(2) When Y is regarded as independent:

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = r \frac{\sigma_x}{\sigma_y}$$

$$a_{xy} = M_X - b_{xy}M_Y$$

When the values obtained in the preceding examples are substituted in the equations for Y as the independent variable,

$$b_{xy} = \frac{28}{70} = 0.683 \frac{1.549}{2.646} = 0.40$$

$$a_{xy} = 6 - (0.4 \times 8) = 2.8$$

Hence, if test scores are predicted on the basis of sales records, the regression equation is

$$T_{xy} \text{ or } X' = a_{xy} + b_{xy}Y = 2.8 + 0.4Y$$

The regression thus defined is charted in Fig. 15·2, where it is contrasted with the regression of Y on X. When this regression equation is used, the predicted test score for sales of 9.17 (thousands of dollars) is found as

$$T_{xy} \text{ or } X' = 2.8 + 0.4\,(9.17) = 6.47$$

The standard error of estimate for this regression may be calculated in a manner similar to that employed for the same purpose in connection with the regression of Y on X, except that the measurement is in X units. Hence

$$\sigma_{d_{xy}} = \sigma_x \sqrt{1 - r^2}$$

$$= 1.549\sqrt{1 - 0.467} = 1.549\sqrt{0.533} = 1.13$$

Coefficients of determination and alienation.—It is some-times convenient to express measures of covariation in two series in terms of percentages. The coefficient of correlation is not satisfactory from this point of view, but its square represents the percentage of total squared variation in the dependent variable that is accounted for by the squared variation in the trend. Thus, by reference to Examples 15·1 and 15·2, it will
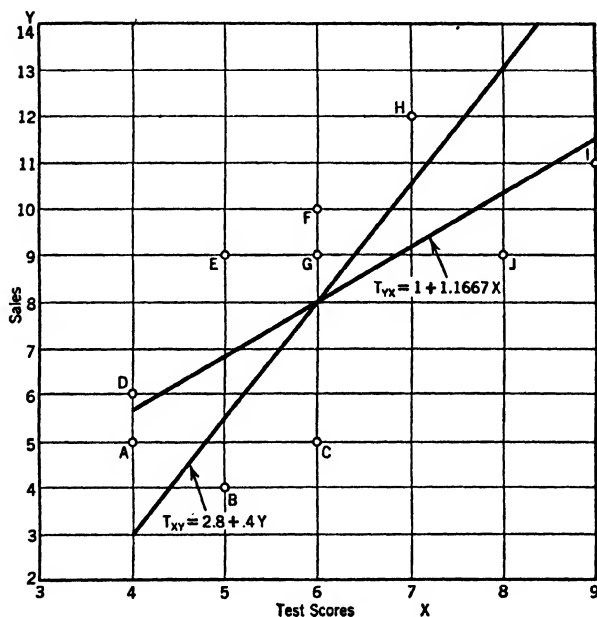
FIG. 15·2.—Regression Lines with Test Scores Independent (T_{YX}) and with Sales Independent (T_{XY}). Data: see Example 15·2.

be seen that the total variance in the trend ($\Sigma t^2 = N\sigma_t^2$) is 32.67, while the total variance in the Y series ($\Sigma y^2 = N\sigma_y^2$) is 70. The ratio of the former to the latter is the square of the coefficient of correlation, that is,

$$r^2 = \frac{32.67}{70} = 0.467$$

The measure of covariation thus obtained is known as the *coefficient of determination* (symbol r^2).

Reference has been made to the errors of estimate (cf.

Example 15·2), and it will be seen that the total variance in this series is 37.3, and the variance or squared standard error of estimate (σ_d^2 or S^2) is 3.73. The ratio of this "unaccounted-for" variance to the variance of the dependent Y series represents a measure of the failure of the series to correlate. It is known as the *coefficient of non-determination* (symbol k^2), and it represents the percentage of variance in the dependent variable that is unaccounted for by the straight-line regression. In the illustrative example,

$$k^2 = \frac{37.3}{70} = 0.533$$

and it will be noted that the sum of these two measures must always be unity, that is,[1]

$$r^2 + k^2 = 1$$

which means that the percentage of variance "accounted for" plus that "unaccounted for" must equal the whole, since, as has been noted, $\Sigma t^2 + \Sigma d^2 = \Sigma y^2$.

The square root of the coefficient of non-determination, k^2, is described as the *coefficient of alienation* (symbol k). Just as r measures the degree of correlation, so k measures the lack of correlation. In the example in question,

$$k = \sqrt{0.533} = 0.73$$

The standard error of r.—Reference has already been made to a simple means of appraising the reliability of the coefficient of correlation. Obviously, a major factor in determining that reliability is the number of paired items available for comparison, since the fewer the items the greater the possibility of accidental

---

[1] It is sometimes assumed that, because r^2 is the proportion of variance explained, it therefore measures the accuracy of prediction. It is clear that r^2 or r might be taken as a measure of comparative predictability. Many statisticians prefer, however, a measure based on the reduction in the error of estimate that is accomplished through use of the regression as a means of prediction. Thus, without the regression, the standard error of estimate of a Y value for a given X value is σ_y. When the regression is used, this error is σ_d. The reduction in error thus effected is $\sigma_y - \sigma_d$. It may be described in percentage terms as $(\sigma_y - \sigma_d)/\sigma_y$, which is identical with $1 - k$ or $1 - \sqrt{1 - r^2}$. In the last form, $1 - \sqrt{1 - r^2}$, this measure of the reduction in error is described as the *index of prediction*.

covariation. The simplest and most satisfactory method of appraising reliability compares given values of r with 5 per cent and 1 per cent chance values, as has been explained.[1] Tables and charts prepared for that purpose are scaled for various values of N or n (cf. Appendix, pages 559–560).

Until recently, however, such calculations of chance values were not available. Under these circumstances, the reliability of r was generally appraised by calculation of its *standard error* or its *probable error*. The standard error of r for large samples is calculated as

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

But the variable skewness of the sampling distribution of r makes this standard error an inexact measure of reliability.

The probable error.—Reference is also often made, in considering the statistical reliability of the coefficient of correlation, to the *probable error* (PE_r), which is that fraction of the standard error which in normal distributions represents a 50 per cent probability. The probable error of r was calculated as 0.6745 times the standard error. But neither the standard error nor the probable error is now much used in evaluating the reliability or r. The greater convenience and more critical appraisal provided by tables or charts such as the table of F have made reference to these older and cruder measures unnecessary.

---

[1] As was suggested in the preceding chapter (page 334), tables or charts of 5 and 1 per cent chance levels of r are related to a much more comprehensive table of the sampling distribution of a statistic called F, or in its original form z. This statistic may be used to measure correlation, with corrections for sampling, by ratios of variances, or mean squares, of regression and estimate, thus

$$F = \frac{\Sigma t^2 \div (m - 1)}{\Sigma d^2 \div (N - m)}$$

where m is the number of series correlated, or constants in the regression equation. Hence, for simple correlation,

$$F = \frac{\Sigma t^2}{\Sigma d^2} (N - 2) = \frac{r^2}{1 - r^2} (N - 2)$$

The 5 and 1 per cent levels of the sampling distribution of F are given in the Appendix, pages 586–589, and a brief discussion of the problem involved appears on pages 554–556.

The crude-data method.—Preceding paragraphs have described the elementary phases of the process of linear correlation. In practice, many short cuts are possible, by means of which much of the detailed manipulation is avoided. In one such practical procedure, steps are taken which make it entirely unnecessary to find the individual deviations from the means. Rather, the sum of the squared deviations, Σx^2 and Σy^2, and the cross products, are secured by manipulating the original X and Y data.

This result is accomplished by use of the correction or " reducing " equations, previously discussed (cf. page 331). Thus the required values of Σx^2, Σy^2, and Σxy are found as:[1]

(1) $\Sigma x^2 = \Sigma X^2 - M_X \Sigma X, \quad \text{or} \quad \Sigma X^2 - \dfrac{(\Sigma X)^2}{N}$

(2) $\Sigma y^2 = \Sigma Y^2 - M_Y \Sigma Y, \quad \text{or} \quad \Sigma Y^2 - \dfrac{(\Sigma Y)^2}{N}$

(3) $\Sigma xy = \Sigma XY - M_X \Sigma Y,$

           $= \Sigma XY - M_Y \Sigma X, \quad \text{or} \quad \Sigma XY - \dfrac{\Sigma X \Sigma Y}{N}$

In Example 15·3, this method is illustrated. It will be noted that squares and products of the crude data are first found. Then these totals are reduced to the more familiar deviation totals, after which the latter are substituted in the usual formulas for the correlation coefficient.[2]

[1] These formulas simply "bunch" tne centering by subtracting, in effect, ΣM_x^2 from ΣX^2, etc. The validity of this method is algebraically proved on page 118. It should be noted that the correction term in (3) is either $M_X \Sigma Y$ or $M_Y \Sigma X$, and that both of these equal $\Sigma X \Sigma Y / N$ or $N M_X M_Y$.

[2] By simple algebraic manipulation, the whole process can, of course, be summarized in a single equation. To simplify computations the numerator and denominator of each fraction are multiplied by N.

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{N \Sigma XY - \Sigma X \Sigma Y}{\sqrt{(N \Sigma X^2 - (\Sigma X)^2)(N \Sigma Y^2 - (\Sigma Y)^2)}}$$

The trend equation may be similarly written, without reference to the deviations. Values of a and b may be found as follows:

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{\Sigma XY - M_X \Sigma Y}{\Sigma X^2 - M_X \Sigma X} = \frac{N \Sigma XY - \Sigma X \Sigma Y}{N \Sigma X^2 - (\Sigma X)^2}$$

$$a = M_y - b M_x = \frac{\Sigma Y - b \Sigma X}{N}$$

Example 15·3

DIRECT CORRELATION OF CRUDE DATA

Data: Scores in psychological tests (X) and sales (Y) of a group of 10 salesmen (see Example 15·2).

Records of salesmen			Squares and products			Regression
Salesmen	X values arrayed	Y values actual	X^2	XY	Y^2	Y values estimated
A	4	5	16	20	25	5.67
D	4	6	16	24	36	5.67
B	5	4	25	20	16	6.83
E	5	9	25	45	81	6.83
C	6	5	36	30	25	8.00
G	6	9	36	54	81	8.00
F	6	10	36	60	100	8.00
H	7	12	49	84	144	9.17
J	8	9	64	72	81	10.33
I	9	11	81	99	121	11.50
Totals........	60	80	384	508	710	80.00
Means........	6	8				
Corrections[1]..			360	480	640	
Centered......			$\Sigma x^2 = 24$	$\Sigma xy = 28$	$\Sigma y^2 = 70$	

[1] Corrections:

$$C_x = \frac{(\Sigma X)^2}{N} = M_X \Sigma X = 6 \times 60 = 360$$

$$C_{xy} = \frac{\Sigma X \Sigma Y}{N} = M_X \Sigma Y \quad \text{or} \quad M_Y \Sigma X = 6 \times 80 \quad \text{or} \quad 8 \times 60 = 480$$

$$C_y = \frac{(\Sigma Y)^2}{N} = M_Y \Sigma Y = 8 \times 80 = 640$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{28}{\sqrt{24 \times 70}} = \frac{28}{40.99} = 0.683$$

Regression:

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{28}{24} = 1.1667$$

$$a = M_Y - b M_X = 8 - 1.1667 \times 6 = 1$$

Correlation of grouped data.—The techniques described in the preceding chapter and the early pages of this chapter are intended for use with ungrouped data. Where data are so extensive as to make grouping desirable, or where they are available only in grouped form, a modification of these techniques is necessary.

Ordinarily, grouped data are classified according to one criterion, as represented by their numerical magnitudes. But
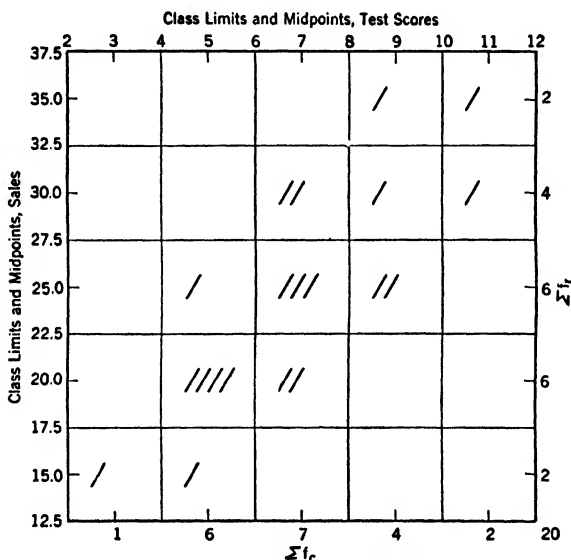

Fig. 15·3.—Simple Bivariate Scatter Diagram. (Data of Table 15·1.

for correlation purposes they must be classified according to each of the two conditions in which covariation is to be measured. Thus, if data to be correlated represent (1) admission-test scores of applicants for positions, and (2) the sales made by those who were accepted for employment, then the accepted salesmen must be classified simultaneously according to (1) their test scores, and (2) their individual sales in dollars. This result is obtained by preparing what is usually described as a scatter diagram or bivariate chart. As in correlation applied to ungrouped data, the scale from left to right represents the

independent variable, here assumed to be test scores. The vertical scale represents the dependent variable, in this case, sales. It is set down in descending order as in a chart, so that the completed scatter diagram may suggest to the eye the nature of the correlation. Class intervals are selected according to the usual principles followed in tabulation.

If the class limits are represented by lines, the scatter diagram will appear as a multi-celled rectangle, such as is illustrated in Fig. 15·3. Each salesman's position in these cells may then be determined and indicated by a slant mark, small circle, or dot. Thus a salesman whose score is 7.7 and whose sales are

TABLE 15·1

DATA TO BE TABULATED AND CORRELATED

Data: Assumed test scores and sales (thousands of dollars):

Sales-men	Test scores	Sales	Sales-men	Test scores	Sales	Sales-men	Test scores	Sales
A	7.7	23	H	6.6	19	O	6.3	25
B	9.0	30	I	8.2	27	P	9.8	23
C	5.2	21	J	11.0	30	Q	4.5	18
D	11.0	35	K	6.4	32	R	7.4	21
E	7.0	27	L	5.0	25	S	5.5	22
F	5.0	15	M	7.6	28	T	4.8	19
G	3.0	15	N	9.0	35			

23 (thousands of dollars) appears as a check in the cell defined by the horizontal class mark 7 and the vertical class mark 25. When all the paired items have been checked in the scatter diagram, the frequency of each cell may be noted by counting the marks in that cell, as shown by the numerals in Example 15·4.

Actual calculation of the coefficient and regression may best be illustrated by reference to a simplified example, in which the number of classes is reduced so that the nature of the manipulations is more readily apparent. The data of Table 15·1 have been prepared for this purpose, and one type of procedure is illustrated in Example 15·4, while another is available in

EXAMPLE 15·4

CORRELATION OF GROUPED DATA

Data: Assumed test scores and sales as given in Table 15·1 and Fig. 15·3.

(1) Double-frequency table (scatter diagram):

Y \ X	3	5	7	9	11	Σf_r
35				1	1	2
30			2	1	1	4
25		1	3	2		6
20		4	2			6
15	1	1				2
Σf_c	1	6	7	4	2	20

(2) Means and variability:

X	f	fX	fX^2	Y	f	fY	fY^2
3	1	3	9	15	2	30	450
5	6	30	150	20	6	120	2,400
7	7	49	343	25	6	150	3,750
9	4	36	324	30	4	120	3,600
11	2	22	242	35	2	70	2,450

Totals   20   140   1,068   20   490   12,650

Means   7   24.5

Corrections for Σx^2 and Σy^2:   980   12,005

Centered squares: $\Sigma x^2 = 88$ $\Sigma y^2 = 645$

(3) Covariability (XY by rows):

XY:	315	385	210	270	330	125	175	225	100	140	45	75
f:	1	1	2	1	1	1	3	2	4	2	1	1
fXY:	315	385	420	270	330	125	525	450	400	280	45	75

$$\Sigma xy = \Sigma XY - M_X \Sigma Y = 3,620 - (7 \times 490) = 190$$

(4) Correlation and regression:

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{190}{\sqrt{88 \times 645}} = 0.7975 \quad (1\% \text{ chance } r = 0.56)$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{190}{88} = 2.1591$$

$$a = M_y - bM_x = 24.5 - 2.1591 \times 7 = 9.3863$$

$$T \text{ or } Y' = a + bX = 9.3863 + 2.1591X$$

Example 15·5. The first form measures correlation and regression by a rather simple method which is similar to that described in connection with the fitting of straight-line trends. The second form utilizes one of the many available " correlation tables " which are characteristic of laboratory practice, particularly if the data are comprehensive. Attention may first be directed to the method illustrated in Example 15·4.

The essential part of the procedure presented in Example 15·4 is the double-frequency table, or scatter diagram shown in section (1). The way in which this table is prepared from original records has already been described (cf. Fig. 15·3). It registers not only the variability of the X and Y series considered separately, but also their covariation. The measurement of the X and Y variabilities just mentioned appears in section (2). The X and Y class marks and frequencies are obtained directly from the double-frequency table, the frequencies being the column and row totals, respectively. The centered squares, Σx^2 and Σy^2, are calculated by familiar methods which do not require comment.

In section (3) covariation as measured by Σxy is computed. The XY products are obtained by reference to each cell of the double-frequency table, taken in succession. For example, the first cell of the first row has an X class mark of 9 and a Y class mark of 35. Hence XY is found as 9×35 or 315 and has a frequency of 1. In the next cell similarly $XY = 11 \times 35 = 385$, with a frequency of 1. In the first cell of the second row $XY = 7 \times 30 = 210$, with a frequency of 2, or an aggregate for this cell of $fXY = 2 \times 210 = 420$. All the fXY's thus obtained are totaled as $\Sigma XY = 3,620$ and are centered in the usual manner. It is thus found that $\Sigma xy = 3,620 - 3,430 = 190$.

In section (4), correlation (r) and regression $(T$ or $Y')$ are computed in the usual way. For $N = 20$ (page 560) the least highly significant r is 0.561. Hence it may be concluded that the results obtained represent a significant rather than a chance covariance.

Attention may next be directed to the correlation table illustrated in Example 15·5. It will be seen that the data are

<div align="center">

EXAMPLE 15·5

CORRELATION AND REGRESSION, GROUPED DATA

</div>

Data: Assumed test scores and sales (see Table 15·1).
Arbitrary origins near the center of the distributions.

Class Limits		2–4	4–6	6–8	8–10	10–12	Calculations for σ_y				Calculations for xy	
	Class Mark	3	5	7	9	11	f	d_y	fd_y	fd_y^2	$\Sigma f_r d_x$	$\Sigma f_r d_x d_y$
32.5–37.5	35				1	1	2	2	4	8	3	6
27.5–32.5	30			2	1	1	4	1	4	4	3	3
22.5–27.5	25		1	3	2		6	0	0	0	1	0
17.5–22.5	20		4	2			6	−1	−6	6	−4	4
12.5–17.5	15	1	1				2	−2	−4	8	−3	6
	f	1	6	7	4	2	20	−2	26		0	19
Calcula-	d_x	−2	−1	0	1	2		Σd_y	Σd_y^2		Σd_x	$\Sigma d_x d_y$
tions for σ_x	fd_x	−2	−6	0	4	4	0	Σd_x				
	fd_x^2	4	6	0	4	8	22	Σd_x^2				
Calcula-	$\Sigma f_c d_y$	−2	−6	0	3	3	−2	Σd_y			Note checks on values of n, Σd_x, Σd_y, and $\Sigma d_x d_y$	
tions for xy	$\Sigma f_c d_y d_x$	4	6	0	3	6	19	$\Sigma d_x d_y$				

$$i_x = 2 \qquad\qquad R_x = 7$$
$$i_y = 5 \qquad\qquad R_y = 25$$

(1) Calculation of the coefficient of correlation, r_{yx} (manipulations in class-interval units):

$$\sigma_x = \sqrt{\frac{\Sigma d_x^2}{N} - \left(\frac{\Sigma d_x}{N}\right)^2} = \sqrt{\frac{22}{20} - \left(\frac{0}{20}\right)^2} = \sqrt{1.1} = 1.04881$$

$$\sigma_y = \sqrt{\frac{\Sigma d_y^2}{N} - \left(\frac{\Sigma d_y}{N}\right)^2} = \sqrt{\frac{26}{20} - \left(\frac{-2}{20}\right)^2} = \sqrt{1.299} = 1.13578$$

$$\Sigma xy = \Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N} = 19 - (0)(-2) = 19$$

$$r_{yx} = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{19}{(20)(1.049)(1.136)} = 0.797$$

(2) Measures of the distributions (changed from class-interval units to original units):

$$M_X = R_x + \frac{\Sigma d_x}{N}\, i_x = 7 + \left(\frac{0}{20}\right)(2) = 7$$

$$M_Y = R_y + \frac{\Sigma d_y}{N}\, i_y = 25 + \left(\frac{-2}{20}\right)5 = 25 - 0.5 = 24.5$$

$$\bar{\sigma}_x = \sigma_x i_x = 1.04881 \times 2 = 2.0976; \bar{\sigma}_y = \sigma_y i_y = 1.13578 \times 5 = 5.67890$$

(3) Calculation of the regression (using original units):

$$b = r\frac{\sigma_y}{\sigma_x} = 0.797\left(\frac{5.6789}{2.0967}\right) = 2.159$$

$$a = M_Y - bM_X = 24.5 - 2.159(7) = 9.387$$

$$T = 9.39 + 2.16X$$

grouped in classes with class marks and frequencies as in Example 15·4. The procedure calculates the standard deviation as described in Chapter VI with the unimportant exceptions that the Y order is the reverse of that previously illustrated, and the X series appears in rows instead of columns. The major steps in the process include: (1) selection of one class mark near the center of each array as an arbitrary origin (R_x and R_y); (2) notation of deviations in class-interval units from each of these arbitrary origins (column d_y and row d_x); and (3) calculation of the standard deviations, σ_x and σ_y, in the manner described in Chapter VI, making use of the columns and rows indicated for this purpose in the example and of the usual correction formulas, as shown in the lower portion of the illustration. Further steps include discovery of the value of Σxy, which is available from the last two columns and from the last two rows (note the check thus provided on this calculation), and substitution of the requisite values in the usual formula for the coefficient.

Up to this point, all calculations may be accomplished in terms of class-interval units, but to obtain the measures of the actual distributions (σ_x and σ_y, and the means) and the regression equation it is necessary to translate these measures into original units by multiplying by class intervals, as shown in the example.

The only step in the procedure that is likely to cause difficulty is the calculation of the $\Sigma f_r d_x$ and the corresponding $\Sigma f_c d_y$. Each item in the column, $\Sigma f_r d_x$, is obtained by multiplying the cell frequencies characteristic of its row by the respective d_x's of the columns in which each appears. For instance, the first $\Sigma f_r d_x$ is calculated as

$$(1 \times 1) + (1 \times 2) = 3$$

The $\Sigma f_c d_y$'s are found in a similar manner. The total of the final column is the sum of products in which each cell frequency in the table is multiplied by its coordinate d_x and d_y, and the total of the last row is the same.

The regression equation for grouped data may be discovered

by the use of the decoded measures, just as in previous examples:[1]

$$T = 9.39 + 2.16X \text{ (original scales)}$$

Corrections for sampling and estimates of reliability may be made as for ungrouped data.

Other grouped data methods.—It should be said that the method of double grouping and the procedure of computation may take many different forms. Printed "checkerboard" tables, with each step in the computation indicated, are available. Some of these forms employ a distribution of the data obtained by adding the frequencies according to the diagonals in the tabulation as an indirect means of obtaining ΣXY (see Appendix, page 538). The essentials of the computation may be carried out on tabulating machines.

In the handling of statistical data such as have been presented, numerous short cuts are possible. If the numbers are large, they may be rounded to perhaps three significant figures, and, in some cases, two figures may suffice to give substantial accuracy. It is also possible to extend the coding process, illustrated in Example 15·5 by utilizing deviations from an arbitrary origin, to cover ungrouped data. In other words, each series may be coded by multiplying or dividing it by a

---

[1] The reduction of the X and Y scales by writing them as unit deviations from arbitrary origins R_x and R_y is a form of coding. Coded scales are often written as X and Y, in which case the original scales may be distinguished as $\bar{X}$ and $\bar{Y}$. It is often convenient, particularly with complex regression equations, to find the regression first for the coded data. Thus computed, $a = -0.1$ and $b = 19/22 = 0.8636$. The regression may then be "decoded" into terms of the original scales (distinguished by bar) by defining X as $(\bar{X} - R_x) \div i_x$ and T or Y' as $(\bar{T} - R_y) \div i_y$, then $T = a + bX$ becomes

$$\frac{\bar{T} - R_y}{i_y} = a + b\left(\frac{\bar{X} - R_x}{i_x}\right)$$

$$\bar{T} = i_y\left[a + b\left(\frac{X - R_x}{i_x}\right)\right] + R_y$$

$$= 5\left[-0.1 + 0.8636\left(\frac{\bar{X} - 7}{2}\right)\right] + 25$$

$$= -0.5 + 2.159\bar{X} - 15.113 + 25$$

$$\bar{T} \text{ or } Y' = 9.39 + 2.16\bar{X}$$

This method is also adaptable to more complex regressions.

constant and expressing it in deviations from an arbitrary origin. The purpose of such coding is simply the reduction of computations. It affects such measures as the means and standard deviations, but not the correlation. Decoding may be carried out as already described.

<div align="center">

EXAMPLE 15·6

RANK CORRELATION

</div>

Data: Test scores and sales (see Example 14·1, page 331).

Sales-men	Test scores Data of X	Sales Data of Y	X Rank	Y Rank	Rank differences d	d^2
A	4	5	9.5	8.5	1.0	1.00
B	5	4	7.5	10.0	−2.5	6.25
C	6	5	5.0	8.5	−3.5	12.25
D	4	6	9.5	7.0	2.5	6.25
E	5	9	7.5	5.0	2.5	6.25
F	6	10	5.0	3.0	2.0	4.00
G	6	9	5.0	5.0	0.0	0.00
H	7	12	3.0	1.0	2.0	4.00
I	9	11	1.0	2.0	−1.0	1.00
J	8	9	2.0	5.0	−3.0	9.00
			55.0	55.0	0.0	$\Sigma = 50.00$

$$r_r = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)} = 1 - \frac{6 \times 50}{10 \times 99}$$

$$= 1 - \frac{300}{990} = 1 - 0.303 = 0.697$$

$$a = (N + 1)(1 - r) + 2 = \frac{11 \times 0.303}{2} = 1.667$$

$$b = r_r = 0.697$$

$$T = 1.667 + 0.697X$$

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}} = \frac{1 - 0.486}{\sqrt{10}} = \frac{0.514}{3.162} = 0.163$$

Method of "rank differences."—In many cases, an adaptation of the Pearsonian method of correlation may be conveniently applied to the rankings of the data instead of to the original items. Data are simply arrayed and ranked from smaller to larger, or vice versa. The substitution of the ranks for the actual data makes the correlation analogous to the calculation of medians and quartiles; that is, each number is given emphasis primarily in accordance with its position in the array rather than in accordance with its specific magnitude. The method is illustrated in Example 15·6. The formula for the coefficient so calculated is

$$r_r = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$
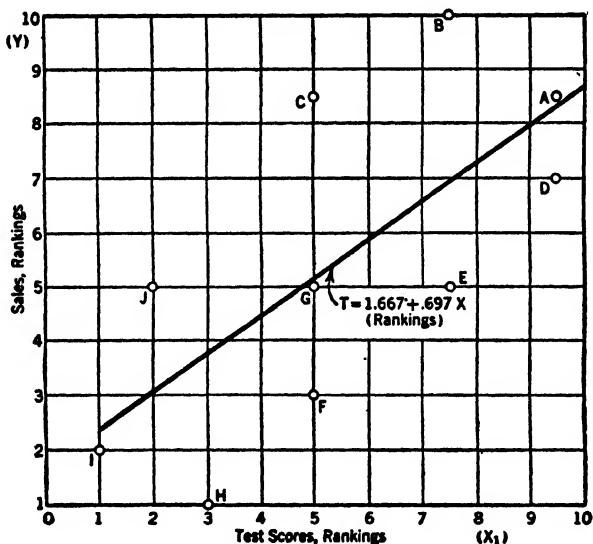
Fig. 15·4.—The Regression Line Based on Rankings of the Data (see Example 15·6).

where d is the difference between the correlative ranks $(X - Y)$ of each case. The regression line passes through the average rank, $(N + 1)/2$, on each scale and has a slope of r_r (see Fig. 15·4). Except for a slight discrepancy arising from the common methods of dealing with ties, if obtained from normally distributed data, a given value of r_r will be reduced by only one or two points from the usual r, and the method may offer actual

advantages in dealing with irregular data. For example, if one correlative item is in a very extreme position in both the X and Y scales, it will ordinarily be given an exceptional weight, because it appears as two squares and a cross product. Ranking removes this emphasis, as in calculating a median. The method is not conveniently applied to large sets of data, however, because of the difficulty of ranking long series of items.

The order of ranking may be from the largest to the smallest or from the smallest to the largest, but reversing the order in only one series will change the sign of the coefficient. When ties occur in the rankings, the tied items may be ranked in the order of their tabulation or by any other chance criterion. Most frequently tied items are designated by the average rank of such items. Thus, if in the ranks 1, 2, 3, and 4 the items ranked 2 and 3 are ties, each would be ranked 2.5. There is no precise method of handling ties, however, and if many of them appear and cannot be resolved by resort to a more exact statement of the data, the method of correlation by rank differences should be avoided.[1]

Fourfold correlation.—It is sometimes important to discover whether data representing two dual classifications are significantly correlated. Such data constitute an abbreviated double-frequency tabulation having only two classes on each axis, as is illustrated in Example 15·7. A linear coefficient (r) might be calculated by the usual method, but a simpler measure ϕ, known

---

[1] A still simpler but rather crude method of correlation, which is sometimes useful in a preliminary investigation, is the method of concurrent deviations. Each deviation in the two correlative series is first labeled plus or minus, with respect to the preceding item, the two adjacent items, the normal, or any other desired basis. These plus and minus series are then correlated by counting the number of agreements ($+$ with $+$, or $-$ with $-$) and the number of disagreements ($+$ with $-$, or $-$ with $+$). If the former number is larger, the correlation is positive; if smaller, it is negative. In either case, the larger sum is labeled C (number of concurrences), and the coefficient is found as follows:

$$\text{Coefficient} = \pm \, [(2C - N) \div N]^{\frac{1}{2}}$$

where N is the number of pairs of correlative items. Neutral concurrences, involving a zero deviation, may be equally divided between the positive and negative concurrences.

EXAMPLE 15·7

FOURFOLD CORRELATION

Data: Double-frequency tabulation of 195 workers, classified as trained or untrained, and as failing or succeeding in the technique for which the training was provided.[1]

$\diagdown$ X Y	Failed 0	Succeeded 1	
Trained 1	$a = 52$	$b = 25$	$77 = g$
Untrained 0	$c = 95$	$d = 23$	$118 = h$
	$e = 147$	$f = 48$	$195 = N$

$$\phi = \frac{bc - ad}{\sqrt{efgh}} = \frac{2375 - 1196}{[(147)(48)(77)(118)]^{\frac{1}{2}}} = 0.147$$

Significance: 5 per cent, $\phi = 0.14$; 1 per cent, $\phi = 0.18$.

as the *coefficient of point correlation*, is more often used. The data may be tabulated as follows:

a	b	g
c	d	h
e	f	N

The measure of correlation is then directly available, as

$$\phi = \frac{bc - ad}{\sqrt{efgh}}$$

where e and f represent the sums of the first and second columns, respectively, and g and h represent the sums of the first and second rows, respectively. In other words,

$$e = a + c \qquad\qquad g = a + b$$
$$f = b + d \qquad\qquad h = c + d$$

---

[1] If the items of the table are small, Yates's correction for continuity may be made. It consists of adding 0.5 to both the smallest item and the item diagonally opposite it, and subtracting 0.5 from the other two items. Column and row totals are thus unchanged. In the example, the corrected items are $a = 52.5$, $b = 24.5$, $c = 94.5$. and $d = 23.5$.

The process of calculation [1] is illustrated in Example $15 \cdot 7$.

Since fourfold correlation gives little meaning to a regression equation and ϕ in itself is of minor importance, its significance may be most conveniently appraised by reference to chi square (which is discussed in some detail on pages 504–509). In fourfold correlation, N times ϕ^2 is equal to chi square. The table of chi square (see page 561) may be used to evaluate significance, and, for fourfold correlation, there is always one degree of freedom $(n = 1)$.

The same result may, of course, be attained directly from the original table, omitting the calculation of ϕ, by measuring chi square as

$$\chi^2 = \frac{N(bc - ad)^2}{e \cdot f \cdot g \cdot h}$$

and it may be noted that the 5 per cent chi-square value for such a table is always 3.84 and the 1 per cent level is always 6.64.

Numerous other special forms of correlation are available for use in connection with less common types of statistical problems.[2]

Improper inferences in correlation.—Throughout the discussion of correlation in these pages, little reference has been made to possible cause-and-effect relationships between or among variables that are correlated. This is not intended to indicate that such a relationship does not exist. As a matter of fact, causal relationship frequently exists, and a knowledge of the facts frequently explains the degree of correlation measured by the various coefficients. Great care must be exercised, however, to avoid misunderstanding the implications of these coefficients. They may represent direct cause-and-effect relationships. They may, in other cases, measure the influence of similar, common, or joint causes affecting both series of variables. They may

---

[1] As thus calculated, without correction for continuity, ϕ is primarily defined as

$$\phi = \sqrt{\frac{a^2}{eg} + \frac{b^2}{fg} + \frac{c^2}{eh} + \frac{d^2}{fh} - 1}$$

[2] See, in this connection, C. C. Peters and W. R. VanVoorhis, *Statistical Procedures and Their Mathematical Bases*, State College, Pennsylvania, Pennsylvania State College, 1935, Chapter X.

and frequently do measure an association in the variables that has little or nothing of causation in it. The fact that such an association is apparent in the coefficients obtained should never, therefore, be accepted in and of itself as demonstrating any direct causal relationship.

It has been found, for instance, that there is an extremely high correlation between the viscosity of asphalt pavements in various weeks throughout the year and infant illness and mortality rates in the same periods. It is probable that there is something in the nature of a mutual causation in which temperature plays a major part. It would obviously be improper, however, to assume anything in the nature of a direct causal relationship, in spite of the high correlation. A causal relationship can be established only through an understanding of the nature of the variables and of the processes involved.

A further caution should be added with respect to the use of regressions in prediction and estimation. It is frequently true that, if a regression is carried beyond the range of original observations, its slope changes. For example, in a given agricultural region, light rains in the growing season may produce light crops, and increasing rainfall may encourage larger crop yields, up to a certain point, beyond which the correlation will gradually but certainly become negative. Regressions of this sort are given special attention in Chapter XVII, but the fact that they are not at all uncommon should suggest the need for caution in the use of strictly rectilinear regressions.

READINGS

(See also special and general references, pages 591 and 597.)

BUTT, W. I., and LOWRY, NELSON, "Education and Size of Family," *Journal of Heredity*, 19 (7), July, 1928, pp. 327–329.

DWYER, PAUL S., and MEACHAM, ALAN D., "The Preparation of Correlation Tables on a Tabulator with Digit Selection," *Journal of the American Statistical Association*, 32 (200), December, 1937, pp. 654–662.

EELS, WALTER CROSBY, "Formulas for Probable Errors of Coefficients of Correlation," *Journal of the American Statistical Association*, 24 (166), June, 1929, pp. 170–173.

EZEKIEL, MORDECAI, "Correlation Coefficients," *American Economic Review*, 19 (2), June, 1929, pp. 246–250.

McINTYRE, FRANCIS, "Automatic Checks on Correlation Analysis," *Journal of the American Statistical Association*, 32 (197), March, 1937, pp. 119–123.

OGBURN, WILLIAM F., "Factors in the Variation of Crime Among Cities," *Journal of the American Statistical Association*, 30 (189), March, 1935, pp. 12–34.

RHODES, E. C., "On the Normal Correlation Function as an Approximation for the Distribution of Paired Drawings," *Journal of the Royal Statistical Society*, 91, 1928 (Part IV), pp. 548–550.

WALLACE, H. A., and SNEDECOR, GEORGE W., *Correlation and Machine Calculation*, Official Publication of Iowa State College of Agriculture and Mechanic Arts, 30 (4), June 24, 1931.

WICKSELL, S. D., "Remarks on Regression," *Annals of Mathematical Statistics*, I (1), February, 1930, pp. 3–14.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Using the data of Exercise 1, page 334, compute the sum of the squared deviations for each series by the centering formula

$$\Sigma x^2 = \Sigma X^2 - M \Sigma X$$

2. Using the data of Exercise 2, page 335, find the regression formulas for each indicated correlation.

3. Using the data of Exercise 3, page 335, find the percentage of variability in the dependent series (X_0) that is explained by each independent series, X_1, X_2, and X_3.

4. Using the data of Exercise 5, page 336, compute coefficients of correlation (each series with X_0) by the ranking method.

5. Calculate σ_x, σ_y, r_{yx}, and T on the basis of the following tabulated data:

(a)

	m's	X scale 2	4	6	8
Y scale	12				1
	9		2	4	1
	6	2	3	1	
	3	1	1		

(b)

$\overline{X}$ $\overline{Y}$		8	10	12	14	16
	X Y	0	1	2	3	4
28	2			1	2	1
24	1		2	4	2	
20	0	1	2	1		

(c)

Y \ X		4	6	8	10	12
Y \ X		0	1	2	3	4
60	4				1	1
50	3			2	1	1
40	2		1	3	2	
30	1		4	2		
20	0	1	1			

ANSWERS TO EXERCISES

1. $\Sigma x_1^2 = 32$; $\Sigma x_2^2 = 72$; $\Sigma x_0 = 200$.

2. For r_{01}: $T = 0.25 + 0.8125X$.
For r_{02}: $T = 2.95 + 0.3525X$.
For r_{03}: $T = 2.74 + 0.6050X$.

3. X_1 explains 54 per cent of X_0 variability.
X_2 explains 53 per cent of X_0 variability.
X_3 explains 77 per cent of X_0 variability.
Each correlation is taken separately.

4. By ranks: $r_{01} = 0.8929$; $r_{02} = 0.8571$; $r_{03} = 0.7143$; $r_{04} = 0.6786$.

5. (a) $\sigma_x = 1.854$; $\sigma_y = 2.364$; $r = 0.716$; $T = 2.973 + 0.914X$.
(b) Decoded: $\sigma_x = 2.000$; $\sigma_y = 2.828$; $r = 0.707$; $T = 12 + X$.
(c) Decoded: $\sigma_x = 2.0976$; $\sigma_y = 11.3578$; $r = 0.7975$; $T = 4.45456 + 4.31818X$.

B. PROBLEMS

6. The data summarized below are assumed to represent the original costs of several machines that were used in a certain manufacturing process, together with the length of service of these machines. It is desired to discover the answers to the following questions:

(a) What machines, classified according to original cost, provided longest service?

(b) What is the measure of association (r) between costs and length of service?

(c) Is the correlation coefficient statistically significant?

Original cost (in thousands of dollars)	Number of units	Length of service (in years)
2.0	2	8
2.0	2	10
2.5	2	9
3.0	3	11
3.2	5	12
4.5	1	13
5.5	3	16
5.6	1	14
6.0	1	15

7. Assume that the following two series represent the weights of a number of parcels and the value of the same items:

Parcel	Weight	Value (in dollars)
A	1	2.00
B	3	1.50
C	4	2.50
D	2	2.00
E	5	3.00
F	8	4.00
G	9	4.00
H	10	5.00
I	13	8.00
J	15	8.00

(a) Prepare a chart showing the covariation in weight and value as it is indicated by these items.

(b) Calculate the coefficient of correlation that measures this covariation.

(c) Check your calculation by securing this coefficient using the raw-data formula.

(d) Fit the linear regression by the method of least squares.

(e) Check the regression just fitted by calculating

$$b = r\frac{\sigma_y}{\sigma_x} \qquad\qquad a = M_Y - bM_X$$

(f) Calculate the standard error of estimate.

(g) Calculate the probable error of the coefficient.

(h) What percentage of the total variance in the value of parcels is apparently accounted for by the variance in weight?

(i) Estimate the probable value of a parcel weighing 12 pounds.

8. Given the following data summarizing the accident experience of the E Company. It compares the number of accidents per employee in a recent year with the number of years of service of the same workers.

Years of Service	Accidents	Years of Service	Accidents
1 1	4 5	6 0	1 8
1 0	2 5	15 0	2 7
1 7	4 0	8 0	1 9
1 6	3 0	10 0	2 0
1 5	3 6	12 0	2 0
1 4	2 6	14 0	2 4
1 3	3 5	26 0	0 9
1 2	2 9	36 0	1 9
1 8	3 1	30 0	1 1
2 0	2 0	34 0	1 5
22 0	3 2	24 0	0 5
20 0	2 2	32 0	1 2
18 0	2 8	28 0	0 9
4 0	1 7	38 0	1 0
16 0	2 7	40 0	1 0

(a) What is the ranking coefficient?

(b) Find the Pearsonian coefficient of correlation.

(c) Chart the relationship.

(d) Estimate the probable accident experience for a worker having 20 years' experience.

9. The following data compare scores made by candidates for civil-service positions (X) with later ratings (Y) of the same individuals. Ratings represent averages of three scores.

X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
78	94	92	82	79	81	100	91	75	78	84	77
67	80	82	75	85	86	81	70	91	90	80	78
75	82	79	77	81	77	89	82	79	76	91	83
76	70	98	81	82	81	97	91	79	78	81	81
92	86	71	86	74	72	72	75	92	84	93	87
66	74	70	56	89	93	79	89	77	79	77	79
99	98	83	89	74	81	84	87	91	93	83	96
99	83	88	94	93	88	94	92	86	82	74	74
78	78	94	94	97	93	93	89	65	74	87	83
94	79	98	97	75	83	81	68	94	85	95	93
58	78	100	100	75	74	86	93	91	82	79	76
85	75	79	82	91	95	71	77	81	67	80	82
88	94	75	77	95	97	87	89	82	73	60	65
80	91	82	86	77	79	71	84	77	76	65	68
80	73	98	93	62	65	88	92	99	90	90	88

(a) Calculate the coefficient of correlation, r.

(b) Chart the relationship of the two measures, and discover the regression equation.

(c) Predict the probable rating for a test score of 91.

10. A chain organization selling automobile supplies recently completed a survey intended to discover how the amounts of average sales in various stores vary with the value of stocks of goods carried in such stores. The results are summarized as follows:

AVERAGE SALES OF 120 STORES CLASSIFIED ACCORDING TO VALUE OF CAPITAL STOCK

Average sale in dollars	Value of capital stock (in thousands of dollars)				
	Up to 4.99	5.00–9.99	10.00–14.99	15.00–19.99	20.00–24.99
Under 0.25	4	2	2	2	1
0.25–0.50	2	3	2	3	2
0.50–0.75	3	4	5	5	1
0.75–1.00	1	3	6	6	3
1.00–1.25	1	3	7	7	5
1.25–1.50	1	2	4	5	8
1.50–1.75	...	1	3	4	5
1.75–2.00	1			1	3

(a) Calculate the coefficient of correlation, r.

(b) Discover the regression equation for the regression of amounts of sales on values of capital stocks.

(c) Estimate the average sale for a store having a capital stock valued at $10,000.

11. The seasonal index for department-stores sales in the United States, for the years 1925 to 1929, is computed in Example 12·1, page 278, using the centered moving-average method. This index was recomputed by the 12-months average method, and both are given below. Determine the degree and the significance of the correlation that exists between the two indexes.

	By Moving Average (1)	By 12-Month Average (2)
J	85 6	85 6
F	83 3	83 4
M	91 7	92 0
A	97 5	97 5
M	99 4	99 4
J	95 0	95 1
J	72 8	72 4
A	76 2	76 2
S	97 5	97 5
O	111 6	111 6
N	117 1	117 1
D	172 3	172 2

12. Management has, for some time, placed major dependence in the selection of certain types of employees on interviewing panels. It compares the decisions of two interviewing panels with respect to 1,000 identical candidates as follows:

INTERVIEWING PANEL I

		Passed	Rejected
INTERVIEWING PANEL II	Passed	73	231
	Rejected	27	669

Use the technique of point correlation to measure the similarity in their decisions, and appraise the significance of this measure.

CHAPTER XVI

MULTIPLE AND PARTIAL CORRELATION

The problem of multiple correlation.—In statistical analysis of business data, and especially in the use of correlation as a means to such analysis, it is unusual indeed to find an association so clear and extensive that all or a large part of the variation in one series is clearly accounted for by change in another. In other words, simple coefficients of correlation sufficiently large so that no further analysis appears necessary and so inclusive that prediction, based upon regression functions, approximates perfection—with errors of estimate reduced to comparative insignificance—are distinctly infrequent. Rather, correlation analysis, in its preliminary stages, is generally of chief value in discovering the limits of such association and in suggesting necessity for further investigation.

In preceding chapters, for instance, the covariation of salesmen's admission-test scores and subsequent sales was measured. The regression, as shown by a fitted trend line, indicates the general nature of the relationship and provides a means for estimating the probable efficiency of prospective salesmen by means of the entrance test. A single test of this nature would probably not result in sufficiently accurate prediction, and the natural question that arises is how the accuracy of prediction may be improved. One possibility suggests itself, namely, the discovery of other conditions similarly related to success, and their combination in a more effective prediction formula. Possibly, for instance, previous business experience may be positively correlated with success, so that prediction based upon both test scores and years of such experience might be somewhat more accurate than that based upon test scores alone.

Sales and experience.—To illustrate the possibilities of such analysis, attention may be directed to the previous employment

experience of the salesmen considered in the example used in the preceding chapter. The facts with respect to this characteristic are summarized in Table 16·1, and in Fig. 16·1 the

TABLE 16·1

RECORDS OF SALESMEN

Data: Assumed for purposes of illustration.

Employees	Psychological test scores X_1	Experience in years X_2	Weekly sales in thousands of dollars Y or X_0
A	4	5	5
B	5	2	4
C	6	4	5
D	4	9	6
E	5	8	9
F	6	4	10
G	6	10	9
H	7	11	12
I	9	10	11
J	8	7	9

regression of sales upon experience is charted. One innovation in both Table 16·1 and Fig. 16·1 should be accorded special attention. It is the designation of the dependent variable, sales, as X_0 instead of Y. In the preceding chapter, where it was desired to relate the process of fitting the regression to that of trend fitting, the dependent variable was designated Y. The change in this chapter is made to facilitate description of various relationships by formulas, in which the subscripts are adequate to indicate the series involved. The standard deviation of the series X_1, for instance, may be simply designated as σ_1 and its mean as M_1, other values being similarly described.

The detailed calculations by means of which the regression of sales on years of experience is fitted are not included here, since they parallel those already explained in Chapter XIV

(see page 340).   The coefficients of correlation and of regression are:

$$r_{02} = 0.696$$

$$b_{02} = 0.628$$

and the regression equation is T or $Y' = 3.605 + 0.628X_2$.

So far as the data indicate, the regression is fairly linear, and salesmen tend to increase their sales with increasing experi-
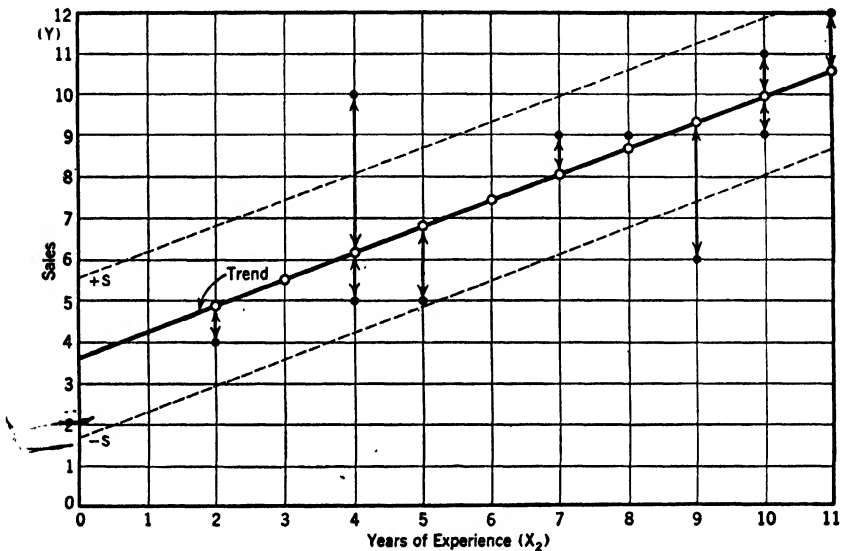
Fig. 16·1.—The Regression of Sales upon Years of Experience.
(Data: See Table 16·1.)

ence, although there are extensive individual variations.   The equation may be taken to mean that, according to this representation of the covariation, salesmen increase their volume of sales by 0.628 (thousands of dollars) for each additional year of experience, beginning with minimum sales of 3.605 (thousands of dollars).

The next problem is to provide a means by which the data with respect to the covariation of sales and test scores may be combined with that relating to the covariation of sales and experience to provide a more effective predictive formula.   This

is essentially the problem multiple correlation analysis presents, i.e., the measurement of relationship between a single dependent variable and a number of independent variables in combination.

The multiple regression.—In the preceding chapter, it was explained that the prediction formula for simple correlation defines a trend line fitted to the data of the two series under consideration. In multiple correlation, the basic problem is the identification of a composite trend line in which two or more independent series are elements. That is, the problem is one of fitting a trend involving three elements: first, a suitable level; second, the data of the first independent series (generally designated X_1); and third, the data of the second independent series (X_2). In the example, these first and second independent variables are test scores and experience, respectively.

The problem, therefore, is essentially the same as that of fitting a parabolic trend, with some slight modification of the elements. For the parabola, the elements are 1, X, and X^2, whereas here they are 1, X_1, and X_2. In the parabola, it is necessary to discover the constants, a, b, and c, measuring the weight given to each element, in order that the total trend may come as close as possible to the data as measured by the criterion of least squares. In the present problem, the same end is sought, but it is customary to use a slightly different notation. The constant associated with the unit element is, as before, indicated by a, but the coefficients associated with X_1 and X_2 are designated b_1 and b_2, respectively. The latter two constants are commonly described as *coefficients of net regression*, and they play a prominent part in the process of partial correlation as well as in that under immediate consideration, that is, multiple correlation.

The normal equations.—The first problem in seeking to combine measurements of covariation among the three variables under consideration is clearly to determine proper weights so that the two known regressions may be expressed in a composite equation. In other words, it is the problem of discovering appropriate values for the two coefficients of net regression, b_1 and b_2. This result is achieved in practically the same manner as the values of the constants for the parabolic trend are deter-

mined, through use of normal equations of the type already discussed in connection with methods of trend fitting. The general form of these equations is required, since X_1 and X_2 take the place of X and X^2, and b_1 and b_2 replace the b and c symbols of the parabola. The required general equations are

(1) $$Na \quad + b_1 \Sigma X_1 \quad + b_2 \Sigma X_2 \quad = \Sigma Y$$

(2) $$a\Sigma X_1 + b_1 \Sigma X_1^2 \quad + b_2 \Sigma X_1 X_2 = \Sigma X_1 Y$$

(3) $$a\Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2 \quad = \Sigma X_2 Y$$

As has been noted, it is convenient to use X_0 for the dependent variable, Y, and the equations are written

(1) $$Na \quad + b_1 \Sigma X_1 \quad + b_2 \Sigma X_2 \quad = \Sigma X_0$$

(2) $$a\Sigma X_1 + b_1 \Sigma X_1^2 \quad + b_2 \Sigma X_1 X_2 = \Sigma X_1 X_0$$

(3) $$a\Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^2 \quad = \Sigma X_2 X_0$$

As in trend fitting, where the equations are abbreviated by time centering, it is possible to reduce the work of calculating the values of b_1 and b_2 by centering each variable, that is, by setting its origin at its mean.[1] In that event ΣX_1, ΣX_2, and ΣX_0 each becomes zero, and equation 1, as well as the first terms in equations 2 and 3, disappear. Then b_1 and b_2 are defined by the equations

(1) $$b_1 \Sigma x_1^2 \quad + b_2 \Sigma x_1 x_2 = \Sigma x_1 y$$

(2) $$b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 \quad = \Sigma x_2 y$$

[1] The normal equations may, of course, be solved without centering. As applied to Example 16·1, they become:

$$10a + 60b_1 + 70b_2 = 80$$
$$60a + 384b_1 + 437b_2 = 508$$
$$70a + 437b_1 + 576b_2 = 614$$

If these are solved algebraically, they provide the same constants as are discovered in Example 16·1. In this case R is found by the formula

$$R^2 = \frac{\Sigma t^2}{\Sigma y^2} = \frac{\Sigma T^2 - M_Y \Sigma Y}{\Sigma Y^2 - M_Y \Sigma Y} = \frac{a\Sigma Y + b_1 \Sigma X_1 Y + b_2 \Sigma X_2 Y - M_Y \Sigma Y}{\Sigma Y^2 - M_Y \Sigma Y}$$

$$= \frac{-0.2704 \times 80 + 0.8394 \times 508 + 0.4620 \times 614 - 8 \times 80}{710 - 8 \times 80} = \frac{48.4512}{70}$$

$$= 0.692.$$

<div align="center">

EXAMPLE 16·1

THE MULTIPLE REGRESSION EQUATION

</div>

Data: See Table 16·1, page 372.

Sales-men	(1) Test scores X_1	(2) Years of expe-rience X_2	(0) Sales in hundreds of dollars X_0	(1)(1) X_1^2	(1)(2) X_1X_2	(1)(0) X_1X_0	(2)(2) X_2^2	(2)(0) X_2X_0	(0)(0) X_0^2
A	4	5	5	16	20	20	25	25	25
B	5	2	4	25	10	20	4	8	16
C	6	4	5	36	24	30	16	20	25
D	4	9	6	16	36	24	81	54	36
E	5	8	9	25	40	45	64	72	81
F	6	4	10	36	24	60	16	40	100
G	6	10	9	36	60	54	100	90	81
H	7	11	12	49	77	84	121	132	144
I	9	10	11	81	90	99	100	110	121
J	8	7	9	64	56	72	49	63	81
Totals	60	70	80	384	437	508	576	614	710
Corrections				360	420	480	490	560	640
Centered squares				24	17	28	86	54	70

Substituting in centered normal equations

(1)	$24b_1 + 17b_2 = 28$
(2)	$17b_1 + 86b_2 = 54$
(1) ÷ 24 : (3)	$b_1 + 0.7083b_2 = 1.1667$
(2) ÷ 17 : (4)	$b_1 + 5.0588b_2 = 3.1765$
	$4.3505b_2 = 2.0098$
	$b_2 = 0.4620$

Substituting in (3)

$$b_1 + (0.7083 \times 0.4620) = 1.1667$$
$$b_1 = 0.8394$$

Then, a is found from the first normal equation as

$$Na = \Sigma X_0 - b_1\Sigma X_1 - b_2\Sigma X_2$$
$$a = M_0 - b_1M_1 - b_2M_2$$
$$= 8 - (0.8394 \times 6) - (0.4620 \times 7) = -0.2704$$

The regression equation is

$$T \text{ or } X_0' = -0.2704 + 0.8394X_1 + 0.4620X_2$$

Or, if X_0 is taken as the symbol of the dependent variable, Y, the equations become

(1) $$b_1 \Sigma x_1^2 \ + b_2 \Sigma x_1 x_2 = \Sigma x_1 x_0$$

(2) $$b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 \ = \Sigma x_2 x_0$$

Finding the constants.—In solving these simultaneous equations, it is frequently convenient, as is indicated in Example 16·1, to manipulate the crude data rather than to calculate the deviations in which the equations are expressed. This is entirely possible, for the required summations may be secured from the results of these manipulations as indicated in the example (see also Example 15·3, page 351).

There are, of course, several alternative methods of solving the simultaneous equations. The procedure followed in the example is selected principally on account of its simplicity and the fact that the student is already familiar with this technique.[1] The Doolittle method, described in the Appendix, page 538, might be used for this purpose.

From the example, it appears that $b_1 = 0.8394$, $b_2 = 0.4620$, and $a = -0.2704$. The multiple regression equation is, therefore,

$$T = -0.2704 + 0.8394X_1 + 0.4620X_2$$

which may be interpreted to mean that sales, beginning with the negative value of a, increase by 0.8394 (thousands of dollars) for each point of advancement in test scores and 0.4620 (thousands of dollars) for each year of experience.

Interpreted in this manner, the regression equation may be employed to predict probable success on the basis of known test scores and experience, just as the simple regression equations were so used in the preceding chapter, and the composite

---

[1] If the centered equations (see above) are solved algebraically for b_1 and b_2 they yield the following formulas, which may be solved as indicated (Np cross products are here used; cf. Example 16·3):

$$b_1 = \frac{\Sigma x_2^2 \Sigma x_1 x_0 - \Sigma x_1 x_2 \Sigma x_2 x_0}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} = \frac{860 \times 280 - (170 \times 540)}{240 \times 860 - 170^2} = 0.8394$$

$$b_2 = \frac{\Sigma x_1^2 \Sigma x_2 x_0 - \Sigma x_1 x_2 \Sigma x_1 x_0}{\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2} = \frac{240 \times 540 - (170 \times 280)}{240 \times 860 - 170^2} = 0.4620$$

relationship may be charted (see Fig. 16·1).[1]   If, for instance, an applicant has a test score of 6 (X_1) and 5 years of experience (X_2), his probable sales record, as estimated on the basis of the relationships thus far discovered, is described by the equation:

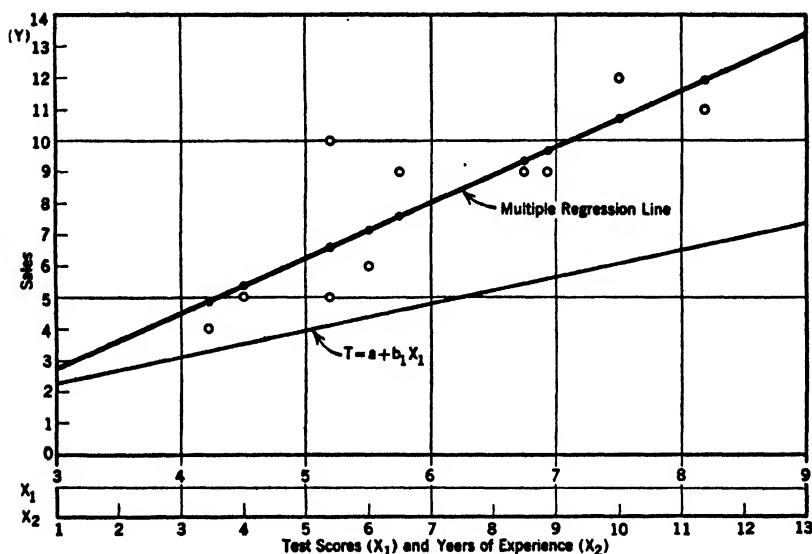$$T = -0.2704 + (0.8394 \times 6) + (0.4620 \times 5) = 7.076$$

FIG. 16·2.—Regression of Sales on Combined Test Scores and Years of Experience (see footnote, below).

Coefficient of multiple correlation.—The most commonly used measure of covariation between two or more independent variables and a dependent series is that known as the *coefficient of multiple correlation* (symbol R), and its meaning is similar

[1] There is little practical value in charts of this type, except as they may be employed to give a graphic estimate.   The X_1 and X_2 scales are preferably united in units of the σ of each, but an approximation which equates the means will serve the purpose. T is found for two or three points on the composite X scale by the regression equation, and the regression line is drawn.   The estimate for each individual salesman is then calculated and plotted on T at the point where T reaches the required height. On this ordinate the actual sales may also be plotted.   Predictions of sales for prospective salesmen may be read from the chart by measuring the difference between the two lines (T and $a + b_1X_1$) on the X_2 ordinate and adding this difference to the lower line ($a + b_1X_1$) on the X_1 ordinate.   The estimated value is read on the X_0 or Y scale.

to that of the coefficient of correlation in that it represents the square root of the ratio of "accounted-for" variance to the total variance to be explained $\left(\dfrac{\Sigma t^2}{\Sigma y^2} \text{ or } \dfrac{\Sigma t^2}{\Sigma x_0^2}\right)$. The multiple coefficient measures the degree of variability in sales that is accounted for by the changes in the combined factors, intelligence and experience. The variance that is "accounted for" is that included in the multiple regression line (T), and the differences between actual sales and the regression line indicate the degree to which intelligence and experience, as thus measured, fail to explain the variability in sales. The standard deviation of these residuals $\left(\sqrt{\dfrac{\Sigma d^2}{N}} = \sigma_d\right)$ is the standard error of estimate. As in linear correlation, it may be readily proved that these three variabilities, measured by either their total squares or variances, are related in this manner:

$$\Sigma y^2 = \Sigma t^2 + \Sigma d^2$$

or,

$$\sigma_y^2 = \sigma_t^2 + \sigma_d^2$$

This relationship is clearly shown in Example 16·2, which compares estimated values with actual values. It will be noted from the example that prediction has been improved by the use of the composite regression equation as compared with the regression on X_1 only, for the "accounted-for" variance has been increased from 3.267 to 4.845; and the "unaccounted-for" variance has been correspondingly decreased from 3.733 to 2.155.

Explained and total variability.—It is obvious that multiple correlation cannot conveniently be measured in terms of the slope of the regression line. It can, however, be measured in terms of the ratio of explained variability to total variability in the dependent series, sales. As in the case of the simple correlation coefficient, where r^2 measured this ratio:

$$R_{0.12}^2 = \frac{\Sigma t^2}{\Sigma y^2} = \frac{\sigma_t^2}{\sigma_0^2}$$

or,

$$R_{0.12}^2 = 1 - \frac{\sigma_d^2}{\sigma_0^2}$$

By means of these equations, the coefficient of multiple correlation may be obtained, as shown in Example 16·2. But in this

<div align="center">

EXAMPLE 16·2

THE MULTIPLE CORRELATION COEFFICIENT

</div>

Data: See Table 16·1.

	(1)	(2)	(0)		
Salesmen	Test scores X_1	Experience X_2	Sales X_0	Estimate T	Error d
A	4	5	5	5.3972	−0.3972
B	5	2	4	4.8506	−0.8506
C	6	4	5	6.6140	−1.6140
D	4	9	6	7.2452	−1.2452
E	5	8	9	7.6226	1.3774
F	6	4	10	6.6140	3.3860
G	6	10	9	9.3860	−0.3860
H	7	11	12	10.6874	1.3126
I	9	10	11	11.9042	−0.9042
J	8	7	9	9.6788	−0.6788
Σ, items	60	70	80	80.0000	0.0000
M	6	7	8	8	
Σ, squares	384	576	710	688.4517	21.5493

Variability.

(1) Aggregate variance to be accounted for:

$$\Sigma x_0^2 = \Sigma X_0^2 - M_0 \Sigma X_0$$
$$= 710 - (8 \times 80) = 70$$

(2) Aggregate variance accounted for:

$$\Sigma t^2 = b_1 \Sigma x_1 x_0 + b_2 \Sigma x_2 x_0$$
$$= (0.8394 \times 28) + (0.4620 \times 54) = 48.451$$

(3) Aggregate variance unaccounted for:

$$\Sigma d^2 = \Sigma x_0^2 - \Sigma t^2$$
$$= 70 - 48.451 = 21.549$$

Correlation.

$$R_{0.12}^2 = \frac{\Sigma t^2}{\Sigma x_0^2} = \frac{48.451}{70} = 0.6921$$

$$R_{0.12} = \sqrt{0.6921} = 0.832$$

example the estimates (T) and the individual errors of the estimate (d) have been shown in detail, merely to explain their nature. In actual practice the regression equation would be applied merely to the records (X_1) and (X_2) of prospective employees. The coefficient of multiple correlation, which is useful as a means of judging the validity of the regression equation, is found by the formulas.[1]

$$\Sigma t^2 = b_1 \Sigma x_1 x_0 + b_2 \Sigma x_2 x_0$$

$$= (0.8394)(28) + (0.4620)(54) = 48.4512$$

$$R^2_{0.12} = \frac{\Sigma t^2}{\Sigma x_0^2} = \frac{48.4512}{70} = 0.6922 \quad \begin{array}{l}\text{(the coefficient of mul-}\\ \text{tiple determination)}\end{array}$$

$$R_{0.12} = \sqrt{0.6922} = 0.832$$

Multiple correlation forms.—In Example $16 \cdot 3$ multiple regression and correlation are combined into a single concise form which calculates the cross products $(P = \Sigma X_1^2, \ \Sigma X_1 X_2,$ etc.) and tabulates them in block P according to the designation in the column headings and stub of the table, as previously explained in connection with trend fitting (cf. page 256). The centered cross products times N $(Np = N \Sigma x_1^2, N \Sigma x_1 x_2,$ etc.) are similarly tabulated in block Np. They are obtained by the usual centering equations

$$N\Sigma x_1^2 = N\Sigma X_1^2 - (\Sigma X_1)^2$$

$$N\Sigma x_1 x_2 = N\Sigma X_1 X_2 - \Sigma X_1 \Sigma X_2, \text{ etc.}$$

The Z items provide a useful running check on the computation. They may be computed like the other numbers, that is, as if Z were merely another X series. But at the same time they should check as the sums of the " full rows," i.e., the rows including the columns in which they begin. Substitutions in

---

[1] Also, $R_{0.12}$ may be obtained from the centered cross products (or the same multiplied by N) by the formula

$$R^2_{0.12} = \frac{\Sigma x_1^2 (\Sigma x_2 x_0)^2 + \Sigma x_2^2 (\Sigma x_1 x_0)^2 - 2\Sigma x_1 x_2 \Sigma x_1 x_0 \Sigma x_2 x_0}{\Sigma x_0^2 [\Sigma x_1^2 \Sigma x_2^2 - (\Sigma x_1 x_2)^2]}$$

$$= \frac{24 \times 54^2 + 86 \times 28^2 - 2 \times 17 \times 28 \times 54}{70(24 \times 86 - 1 \cdot 7^2)} = 0.692$$

the centered normal equations may be made from the Np block. The factor N need not be removed from these substituted items since it affects equally both sides of the usual equations, and both terms of the usual ratios. In most problems it is useful in that decimals are avoided by retaining it.

The form of solution indicated in Example 16·3 will be found particularly adapted to extended problems. In such cases the solution of the normal equations may be carried out by an abbreviated process known as the Doolittle method. This method will be found illustrated in the Appendix, where the same data together with one further independent series are utilized (see page 540).

The Betas.—It has been seen that the b's are in effect weights, expressing the force to be given to the component elements (X_1 and X_2) in combining them into an approximation of X_0. But the units of X_1 may be very different in size and variability from those of X_2 or from other elements that may be employed, and then the b's do not indicate the comparative importance of each series. Hence, the correlation is sometimes calculated by means of normal equations which assume theoretically that each series of the data has been expressed in units of its own standard deviation. The normal equations then become

$$\beta_1 + \beta_2 r_{12} = r_{10}$$

$$\beta_1 r_{12} + \beta_2 = r_{20}$$

where the subscripts 1, 2, and 0 indicate X_1, X_2, and X_0 (or Y), respectively, and where the constants are written as β instead of b in order to distinguish their origin. After the required r's have been found, as explained in the last chapter, the normal equations may be solved, and, for the data of Example 16·2, the beta coefficients are found to be

$$\beta_1 = 0.49152$$

$$\beta_2 = 0.51205$$

The advantage of these constants is the fact that, as contrasted with the b's, they provide a more accurate idea of the relative importance of the elements, X_1 and X_2, in predicting sales,

EXAMPLE 16·3.

MULTIPLE REGRESSION AND CORRELATION

Data: See Examples 16·1 and 16·2.

	Salesmen	Test scores X_1	Experience X_2	Sales X_0	Row Σ Z
	A	4	5	5	14
	B	5	2	4	11
	C	6	4	5	15
	D	4	9	6	19
	E	5	8	9	22
	F	6	4	10	20
	G	6	10	9	25
	H	7	11	12	30
	I	9	10	11	30
	J	8	7	9	24
	Σ	60	70	80	210
Cross products (P).	1	384	437	508	1,329
	2		576	614	1,627
	0			710	1,832
	Z				4,788 ch'k
Same, centered times N (Np)	1	240	170	280	690
	2		860	540	1,570
	0			700	1,520
	z				3,780 ch'k

Substituting in centered equations:

$$240\, b_1 + 170\, b_2 = 280$$

$$170\, b_1 + 860\, b_2 = 540$$

Solving, and substituting in first normal equation (cf. Example 16·1):

$$T \text{ or } Y' = -\,0.2704 + 0.8394 X_1 + 0.4620 X_2$$

Correlation:

$$R^2_{0.12} = \frac{\Sigma t^2}{\Sigma y^2} = \frac{b_1 \Sigma x_1 x_0 + b_2 \Sigma x_2 x_0}{\Sigma x_0^2}$$

$$= \frac{0.8394 \times 280 + 0.4620 \times 540}{700} = 0.692$$

since the dissimilarity of the units in which each was expressed has been eliminated. Moreover, their statistical reliability may be conveniently measured.[1]

The β's may readily be found from the b's rather than by direct calculation, by noting the ratios

$$b_1 \frac{\sigma_1}{\sigma_0} = \beta_1 \quad \text{and} \quad b_2 \frac{\sigma_2}{\sigma_0} = \beta_2, \text{ etc.}$$

or

$$0.8394 \times \frac{1.5492}{2.6458} = 0.4915 \quad \text{and} \quad 0.4620 \times \frac{2.9326}{2.6458} = 0.5121$$

Conversely, if the β's have been calculated directly, the b's may be readily secured for use in the multiple regression equation by reversing these ratios, so that [2]

$$b_1 = \beta_1 \frac{\sigma_0}{\sigma_1} \quad \text{and} \quad b_2 = \beta_2 \frac{\sigma_0}{\sigma_2}, \text{ etc.}$$

[1] The statistical reliability of the two β's may be found as follows (for procedure where three or more β's are involved, see Appendix, page 542):

$$\sigma_\beta = \left[\frac{1 - R^2}{(1 - r_{12}^2)(N - m)} \right]^{\frac{1}{2}} = \left[\frac{0.3078}{0.8600 \times 7} \right]^{\frac{1}{2}} = 0.2261$$

$$t_1 = \frac{\beta_1}{\sigma_\beta} = \frac{0.4915}{0.2261} = 2.17$$

$$t_2 = \frac{\beta_2}{\sigma_\beta} = \frac{0.5121}{0.2261} = 2.26$$

Reference to pages 586–589 indicates that both t_1 and t_2 are slightly below the least significant value, which is 2.365 for $N - m = 7$.

The convenience of these coefficients as a means of determining which series may be regarded as comparatively unimportant in multiple correlation analysis has been clearly indicated by Frederick V. Waugh in his "A Simplified Method of Determining Multiple Regression Constants," *Journal of the American Statistical Association*, 30 (192), December, 1935, pp. 694–700. He concludes that "If any of the regression coefficients are of a magnitude less than twice their standard error, they are non-significant."

[2] A measure of reliability is available with which to judge the significance of the net regression coefficients or b's. Their standard error is defined by the equation, (subscript 3 may be dropped or 4, etc., added as required; if σ_0/σ_1 is dropped, σ_β is indicated),

$$\sigma_{b_{01.22}} = \frac{\sigma_0}{\sigma_1} \left[\frac{1 - R_{0.123}^2}{(N - m)(1 - R_{1.23}^2)} \right]^{\frac{1}{2}}$$

See Mordecai Ezekiel, *Methods of Correlation Analysis*, New York, John Wiley & Sons, 1930, p. 258.

It should be added that, if the β's have been calculated directly by means of the normal equations already described, R may be calculated directly by means of an equation in which the β's appear as weights applied to the simple coefficients of correlation, as follows: [1]

$$R_{0.12}^2 = \beta_1 r_{01} + \beta_2 r_{02} = (0.4915)(0.6831) + (0.5121)(0.6960)$$
$$= 0.692$$
$$R_{0.12} = \sqrt{0.692} = 0.832$$

This formula is merely an adaptation of the one previously used where

$$R_{0.12}^2 = (b_1 \Sigma x_1 x_0 + b_2 \Sigma x_2 x_0) \div \Sigma x_0^2$$

since, when the data are expressed in σ units, b becomes β, $\Sigma x_1 x_0$ becomes $N r_{01}$, $\Sigma x_2 x_0$ becomes $N r_{02}$, and Σx_0^2 becomes N. For more complex coefficients, the formula adds $\beta_3 r_{03}$, etc.

Reliability of R.—When appropriate tables of significanc*y* are available, they provide a more satisfactory method of es ,t-mating the reliability of the coefficient of multiple correlation and the multiple regression equation built about it. The tables describe the coefficients that may be expected to appear by

---

[1] In connection with multiple correlation, it is often desirable to measure the simple correlations existing among the various series studied. These may readily be computed from the Np block of Example 16·3 by noting that the diagonal line (240; 860; 700) represents the centered squares (Σx_1^2; Σx_2^2; Σx_0^2) and the other items, omitting the check column, Z, are centered cross products ($\Sigma x_1 x_2$; $\Sigma x_1 x_0$; $\Sigma x_2 x_0$). A convenient form for the computation is as follows (cf. Appendix, p. 537):

Np block with duplicates (Z's are row sums):

	(1)	(2)	(0)	(Z)
(1)	240	170	280	690
(2)	170	860	540	1570
(0)	280	540	700	1520

Divide each row by the centered squares to get b's.

	(1)	(2)	(0)	(Z)
(1)	1.C00	0.708	1.165	2.873 ch'k
(2)	0.198	1.000	0.629	1.827 ch'k
(0)	0.400	0.771	1.000	2.171 ch'k

Multiply the complementary b's ($b_{21} \times b_{12}$, etc.).

$$r_{12}^2 = 0.708 \times 0.198 = 0.140$$
$$r_{10}^2 = 1.165 \times 0.400 = 0.466$$
$$r_{20}^2 = 0.629 \times 0.771 = 0.485$$

pure chance among samples from uncorrelated data once in 20 times (5 per cent table) and once in 100 times (1 per cent table).

If reference is made to the two charts on pages 559 and 560 in connection with the coefficient discovered in preceding paragraphs, it will be found that, for 10 items and 3 variables, a coefficient of multiple correlation of approximately 0.76 is required to reduce chance to the 5 per cent margin. For 1 per cent, the value is 0.86. Evidently, the coefficient here calculated, 0.832, is well above the lower limit that is established by the table as representing a significant coefficient. It should be concluded, therefore, that the measure of relationship between test scores, experience, and sales is significant, though it may well be supplemented by recourse to additional data.[1]

The standard error of estimate.—It is convenient to have some measure of the statistical reliability of the multiple regression equation in addition to that of the coefficient of multiple correlation. Such a measure is available in the *standard error of estimate* of the multiple regression. This standard error of estimate may readily be calculated (see Example 16·2) in accordance with the formulas

$$\Sigma d^2 = \Sigma y^2 - \Sigma t^2$$

and

$$S^2 \quad \text{or} \quad \sigma_d^2 = \sigma_y^2 - \sigma_t^2$$

and the value may be calculated as

$$\sigma_d^2 = \frac{\Sigma d^2}{N} = 21.55 \div 10 = 2.155$$

or

$$\sigma_d = 1.468 \quad .$$

---

[1] As in simple correlation, an appraisal of reliability is obtainable also by use of the table of F. The statistic F is here another measure of correlation, and its sampling distribution is indicated at the required levels in the table of F. For multiple correlation, the required formula is

$$F = \frac{R^2}{1 - R^2} \times \frac{N - m}{m - 1} = \frac{0.692}{0.308} \times \frac{10 - 3}{3 - 1} = 7.86$$

where m is the number of constants in the regression equation, or the number of series correlated (see pages 586–589). The table of F indicates that the least significant F is 4.74 and the least highly significant F is 9.55.

It will be seen that this calculation parallels that which is used in simple linear correlation. It represents merely the standard deviation of the residuals, $Y - T$, which, in turn, is a measure of the degree to which the regression line fails to conform to the data.

The standard error of estimate thus found may be corrected for the error of sampling by the formula

$$\bar{S}^2 = \sigma_d^2 \times \frac{N}{N - m} = (1.468)^2 \times \frac{10}{10 - 3} = 3.079$$

$$\bar{S} = \sqrt{3.079} = 1.755$$

where m is the number of constants in the regression equation, or the number of series correlated.

In actual practice, it is seldom advisable to calculate individual estimates and errors of estimate, so that neither Σd^2 nor Σt^2 may be readily available. In such cases, the standard error of estimate of the multiple regression for X_1, X_2, and X_0 (corrected for sampling) is more readily calculated as

$$\bar{S}^2 = \frac{\Sigma x_0^2 - \Sigma t^2}{N - m} = \frac{\Sigma x_0^2 - (b_1 \Sigma x_0 x_1 + b_2 \Sigma x_0 x_2)}{N - m}$$

which may be expanded, when necessary, by the addition of terms $(b_3 \Sigma x_0 x_3$, etc.) within the parenthesis. Appropriate values are readily available in Example 16·3.

The value of $\bar{S}$ may be found as

$$\bar{S}^2 = \frac{70 - [(0.8394)(28) + (0.4620)(54)]}{10 - 3} = 3.078$$

$$\bar{S} = \sqrt{3.078} = 1.754$$

Partial correlation.—It may be demonstrated (see Appendix, page 546) that the constants b_1, b_2, etc., featuring the multiple regression equation, represent the net regressions of the dependent variable upon the respective independent series after the effects or influences of the other independent variables have been theoretically removed or held constant. In other words, b_1 is the measure of net regression of X_0 on X_1, and b_2 is the

measure of the net regression of X_0 on X_2, and the idea of its being a *net* regression involves the assumption that its use holds the influence of the other independent series constant. It is for this reason that the b's are referred to as the *net regression coefficients*.

This characteristic of the b's suggests an additional type of correlation analysis, known as *partial correlation*. It proposes to evaluate the relationship between a dependent series and a single independent series when other factors are held constant. It will be recognized at once that this objective is distinctly different from that of simple correlation, where covariation between two series is measured while covariation with all other series is ignored. The objectives of partial correlation are somewhat analogous to those of laboratory experiments in the physical sciences in which it is desired to eliminate or hold constant the influence of a number of factors in order to measure more accurately one particular type of relationship. Partial correlation has, for this reason, a wide field of possible application. It may, for instance, be desired to measure covariation between prices of farm products and known supplies of these products, when the influence of changes in general business activity is held constant.

To achieve these results, the net regression coefficients or b's already calculated are used as measures of the net effect or association of each of the independent series with the dependent series in estimating or predicting values for the dependent variable. Actually, if a detailed analysis is made in which the original measures of correlation between each independent series and the dependent variable are "corrected" for the measured "influence" of the other independent series, the resulting coefficients are identical with those already designated as net regression coefficients. Moreover, it may easily be shown that the beta coefficients (β's) are similar net regression coefficients in which the data are expressed in standard deviation units. For this reason, either the b's or the β's provide a readily available means of making estimates of the covariation in two series while the known relationships with other series are taken into account or held constant.

The process may be illustrated by reference to the data of the example used in preceding sections of this chapter. The net regression coefficient describing the regression of sales on test scores has been found to be $b_1 = 0.8394$. An estimating equation may be readily set up by discovering the value of the constant a as

$$a = M_0 - b_1 M_1 = 8 - (0.8394)(6) = 2.964$$

The estimating equation is, therefore,

$$T = 2.964 + 0.839 X_1$$

where T represents estimated sales, assuming average experience.

Estimates may be made as usual by substituting appropriate test scores for X_1. Similarly, since $b_2 = 0.4620$, the regression equation for sales on experience may be defined by discovering $a = M_0 - b_2 M_2$, so that this regression is defined as

$$T = 4.766 + 0.462 X_2$$

where T represents an estimate of sales, assuming average test scores.

Coefficient of partial correlation.—The usual measure of net covariation defined in such equations as these is the coefficient of partial correlation. It is symbolized with the r characteristic of simple correlation coefficients, but there is always attached a subscript indicating the series being correlated and those that are held constant. The portion of the subscript before the period in it designates the correlated series; the variables following the period are those whose influence is theoretically held constant. Thus $r_{01.2}$ means that the series designated as X_0 and X_1 are being correlated and that the series X_2 is held constant. Similarly, $r_{01.234}$ merely indicates that the influence of two additional series is being isolated in order to measure the net covariation in variables X_0 and X_1.

Calculation of the coefficient of net or partial correlation might proceed from that of the coefficient of net regression, but it is generally more convenient to make use of a technique that combines simple correlation coefficients to determine the partial measure. Thus the measure of partial correlation featuring the dependent series X_0 (symbol 0), the independent series X_1

(symbol 1), and another independent variable, X_2 (symbol 2), which is held constant, is expressed by the equation [1]

$$r_{01.2} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{(1 - r_{02}^2)(1 - r_{12}^2)}}$$

A similar equation expressing the relationship between other independent series and a dependent variable may be constructed from the same pattern by adjusting the subscripts.[2]

When this measure of covariation is applied to the data of the example, where sales are compared with test scores and experience, and where the simple coefficients of correlation are $r_{01} = 0.6831$, $r_{02} = 0.6960$, and $r_{12} = 0.3742$, the covariation of test scores and sales, when experience is held constant, is calculated as

$$r_{01.2} = \frac{(0.6831) - (0.6960)(0.3742)}{\sqrt{(1 - 0.6960^2)(1 - 0.3742^2)}} = 0.635$$

This coefficient represents the correlation of sales and test scores for salesmen who are assumed to have had the same number of years' experience. Similarly, the covariation of sales and years of experience, when test scores are held constant, is

$$r_{02.1} = \frac{(0.6960) - (0.6831)(0.3742)}{\sqrt{(1 - 0.6831^2)(1 - 0.3742^2)}} = 0.650$$

This coefficient represents the correlation of sales with years' experience for salesmen assumed to have equal scores in the psychological test.[3]

---

[1] In general terms, where the subscripts 1, 2, and 3 merely mean the first, second, and third series employed in a particular partial correlation, the formula reads

$$r_{12.3} = (r_{12} - r_{13}r_{23}) \div [(1 - r_{13}^2)(1 - r_{23}^2)]^{\frac{1}{2}}$$

[2] The formula utilizing the r's may for some purposes be more conveniently expressed in terms of the centered cross products or the Np block (cf. Examples 16·1 or 16·3) as follows:

$$r_{01.2} = \frac{\Sigma x_2^2 \Sigma x_1 x_0 - \Sigma x_2 x_0 \Sigma x_1 x_2}{[(\Sigma x_1^2 \Sigma x_2^2 - \overline{\Sigma x_1 x_2^2})(\Sigma x_2^2 \Sigma x_0^2 - \overline{\Sigma x_2 x_0^2})]^{\frac{1}{2}}}$$

$$= \frac{86 \times 28 - 54 \times 17}{[(24 \times 86 - 17^2)(86 \times 70 - 54^2)]^{\frac{1}{2}}} = 0.635$$

[3] Another and rather crude type of correlation analysis is sometimes useful in approximating the results of partial correlation. It is known as *part correlation*, and it would evaluate, in this instance, the correlation of sales and psychological tests with the effects of experience eliminated from the sales data only, and not from that of the psychological tests. In this case, the coefficient of part correlation thus derived

The coefficients of partial correlation have a definitely measurable significance with respect to the proportion of the total variance in the dependent or Y series that they account for, a significance similar to that noted in connection with coefficients of simple and multiple correlation. It will be recalled that the square of the simple coefficient of correlation, known as the coefficient of determination, represents the percentage of total variance in the dependent variable that is accounted for by the variance of the regression line. It is the ratio of variance in T to variance in Y. Similarly the squared coefficient of multiple correlation represents the percentage of total variance that is accounted for by the multiple regression. In so far as the additional series included in the multiple regression results in a reduction in "unaccounted-for" variance, this reduction, expressed as a percentage of the total of "unaccounted-for" variance remaining before the inclusion of any given series in the multiple regression, is the square of the coefficient of partial correlation for that given series. The squared coefficient of partial correlation, therefore, measures the percentage reduction in variance that is attributable to covariation with the particular independent series the coefficient represents. This relationship is shown in Example 16·4.

<div align="center">

EXAMPLE 16·4

THE MEANING OF PARTIAL CORRELATION
</div>

Data: Sales (X_0), test scores (X_1), and years of experience (X_2), as in Examples 16·1 and 16·2.

<div align="center">The meaning of $r_{01.2}$</div>

(a) Total variance to be accounted for (Σx_0^2) as a percentage $= 100.0\%$

(b) Variance accounted for by the simple $r_{02} = 0.696 : r_{02}^2 \quad = 48.4\%$

(c) Remainder unaccounted for, $100 - 48.4 \qquad = 51.6\%$

(d) Variance accounted for by multiple $R_{0.12} = 0.832 : R_{0.12}^2 \; = 69.2\%$

(e) Remainder unaccounted for, $100 - 69.2 \qquad = 30.8\%$

(f) Reduction accounted for by including r_{01} in the Multiple
  $R : 51.6 - 30.8 \qquad = 20.8\%$

(g) Percentage reduction in unaccounted for variance: $\dfrac{20.8}{51.6} = 40.3\%$

(h) Since 40.3% or 0.403 is $r_{01.2}^2$, $r_{01.2} = \sqrt{0.403} = 0.635$

is 0.739. For an adequate discussion of this device, see Mordecai Ezekiel, *Methods of Correlation Analysis*, New York, John Wiley & Sons, 1930, pp. 181–183.

Uses of partial correlation.—The technique of partial correlation may be applied to an indefinite number of independent variables. Thus, in the illustrative example, it might be desired to measure covariation between sales and ages, years of education, intelligence quotients, ratings derived from rating scales, sales in the first month of service, high-school or college grades, and numerous other characteristics. When the partial coefficient refers to but two independent and one dependent series, it is described as of the first order. When three independent variables (X_1, X_2, and X_3) are included, it is said to be of the second order, and the designation varies accordingly with each additional independent variable. In general, coefficients of the second order may be found indirectly by multiple correlation by the use of Rietz's formula (where Y and X_1—subscripts 0 and 1—are correlated, and the influences of other X's are excluded):

$$1 - r_{01.23}^2 = \frac{1 - R_{0.123}^2}{\cdot 1 - R_{0.23}^2}$$

This formula may be expanded to include other independent series by an obvious extension of the subscripts. It may be applied so as to exclude the influence of any selected independents merely by renumbering the series so that the two to be correlated are designated as *0* and *1*.

READINGS

(See also special and general references, pages 591 and 597.)

BAKST, AARON, "A Modification of the Computation of the Multiple Correlation and Regression Coefficients by the Tolley and Ezekiel Method," *Journal of Educational Psychology*, 22 (8), November, 1931, pp. 629–635.

COWAN, DONALD R. G., "A Note on the Coefficient of Part Correlation of a Dependent Variable with All but One of a Group of Other Variables," *Journal of the American Statistical Association*, 27 (178), June, 1932, pp. 177–182.

COURT, ANDREW T., "Measuring Joint Causation," *Journal of the American Statistical Association*, 25 (171), September, 1930, pp. 245–254.

FISHER, R. A., "The General Sampling Distribution of the Multiple Correlation Coefficient," *Proceedings of the Royal Society*, December 3, 1928, pp. 654–673.

FRANZEN, RAYMOND, and DERRYBERRY, MAHEW, "The Routine Computation of Partial and Multiple Correlation," *Journal of Educational Psychology*, 22 (9), December, 1931, pp. 641–651.

GRIFFEN, HAROLD D., "On the Coefficient of Part Correlation," *Journal of the American Statistical Association*, 27 (179), September, 1932, pp. 298–301.

HAMILTON, C. HORACE, "Algebraic Derivation of the Normal Equations Involved in Multiple and Partial Correlation," *Journal of the American Statistical Association*, 28 (182), June, 1933, pp. 204–206.

HEFLEBOWER, R. B., "Factors Relating to the Price of Idaho Potatoes," Idaho Agricultural Experiment Station *Bulletin* 166, 1929.

HORST, PAUL, "A Short Method for Solving for a Coefficient of Multiple Correlation," *Annals of Mathematical Statistics*, 3 (1), February, 1932, pp. 40–44.

KELLY, TRUMAN LAUD, and McNEMAR, QUINN, "Doolittle Versus the Kelly-Salisbury Iteration Method for Computing Multiple Regression Coefficients," *Journal of the American Statistical Association*, 24 (166), June, 1929, pp. 164–169.

MINER, JOHN RICE, "The Standard Error of a Multiple Regression Equation," *Annals of Mathematical Statistics*, 2 (3), August, 1931, pp. 320–323.

SMITH, BRADFORD BIXBY, "Another Attempt to Explain Multiple Correlation in Simple Terms," *Journal of the American Statistical Association*, 24 (165), March, 1929, pp. 61–65.

STOUFFER, SAMUEL A., "Evaluating the Effect of Inadequately Measured Variables in Partial Correlation Analysis," *Journal of the American Statistical Association*, 31 (194), June, 1936, pp. 348–360.

WHERRY, R. J., "A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation," *Annals of Mathematical Statistics*, 2 (4), November, 1931, pp. 440–457.

WILKS, S. S., "On the Sampling Distribution of the Multiple Correlation Coefficient," *Annals of Mathematical Statistics*, 3 (3), August, 1932, pp. 196–203.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Analyze the data summarized below to discover the coefficient of multiple correlation, the multiple regression equation, the coefficient of multiple determination, and the two beta coefficients:

X_1	X_2	X_0 or Y
2	4	9
5	5	12
4	5	10
3	3	7
6	8	17
20	25	55

2. Using the following data, and assuming the X_0 series to be dependent:

(a) Secure values for the multiple regression equation.

(b) Secure the measure of multiple correlation.

(c) What values of the multiple correlation coefficient would be required for 1 per cent and 5 per cent significance, respectively?

(d) Predict the value of the dependent series when $X_1 = 8$ and $X_2 = 8$.

X_1	X_2	X_0
10	10	18
6	5	7
9	7	8
10	10	10
12	12	18
13	14	19
11	9	9
9	5	7

3. Calculate the coefficients of multiple correlation, $R_{1.23}$, $R_{2.13}$, $R_{3.12}$, and $R_{0.123}$ for the following series of data. Also calculate the regression equation, $T_{0.123}$. Measure the betas and test their significance.

X_1	X_2	X_3	X_0
14	20	18	13
6	6	7	6
10	11	8	8
12	28	10	11
14	31	18	13
20	32	19	14
12	25	.9	13
8	7	7	2

4. Making use of the data of the following four series:

(*a*) Find the values of $R_{0.12}$, $R_{0.123}$.

(*b*) Determine the statistical significance of each of the multiple coefficients by two methods.

(*c*) Calculate the partial coefficients, $r_{01.2}$, $r_{02.1}$, $r_{03.1}$, and $r_{03.12}$.

(*d*) Measure the statistical significance of each of the partial coefficients.

X_1	X_2	X_3	X_0
10	25	11	26
6	11	16	15
9	16	14	16
10	33	11	18
12	36	9	26
13	37	7	27
11	30	12	17
9	12	16	15

ANSWERS

1. $R_{y.12} = 0.986$, $T = 0.85 + 0.35X_1 + 1.75X_2$, $\beta_1 = 0.145$; $\beta_2 = 0.860$.

2. (*a*) $T = 1.5888 - 0.5868X_1 + 1.8088X_2$. (*b*) $R_{0.12} = 0.88$.

(*c*) 5 per cent, $R = 0.84$; 1 per cent; $R = 0.92$.

(*d*) 11.36.

3. $R_{1.23} = 0.92$; $R_{2.13} = 0.85$; $R_{3.12} = 0.86$; $R_{0.123} = 0.898$.

$T = 1.9556 + 0.0169X_1 + 0.2776X_2 + 0.1909X_3$; $\beta_1 = 0.0168$, $\beta_2 = 0.6942$, $\beta_3 = 0.2384$; $t_1 = 0.03$, $t_2 = 1.66$, $t_3 = 0.55$.

4. (*a*) $R_{0.12} = 0.755$. (*c*) $r_{01.2} = 0.304$.

$R_{0.123} = 0.931$. $r_{02.1} = 0.235$.

$r_{03.1} = -0.713$.

$r_{03.12} = -0.830$.

5. The following data are assumed to summarize the results of a study of a number of electrical devices examined by a testing laboratory:

Item	A Length of service in 1,000 hours	B Current consumption in watts	C Power developed in arbitrary units
1	1.5	50	2
2	2.5	55	12
3	1.7	57	5
4	1.8	58	3
5	3.0	60	7
6	1.6	62	4
7	1.9	65	5
8	2.0	66	8
9	4.0	67	18
10	2.1	69	3
11	2.4	70	1
12	2.6	72	3
13	3.0	74	4
14	5.0	76	9
15	3.2	78	8
16	3.4	80	6
17	3.6	82	5
18	3.2	84	5
19	6.0	86	3
20	4.2	88	16
21	2.7	90	13
22	4.5	92	11
23	4.9	93	3
24	4.2	96	15
25	5.0	97	14

Find: (a) Coefficients of correlation: r_{ab}, r_{ac}, r_{bc}

(b) The multiple regression equation (A as Y, B as X_1 and C as X_2)

(c) The coefficient of multiple correlation, $R_{a \cdot bc}$

(d) The partial correlation coefficients, $r_{ab \cdot c}$ and $r_{ac \cdot b}$

6. The following summary presents significant facts with respect to a number of samples of cord examined by the testing laboratory of a large wholesale organization:

Sample	Y Tensile strength in pounds	X_1 Number of strands	X_2 Weight in ounces
A	4	75	30
B	10	90	30
C	8	80	35
D	7	85	40
E	6	70	50
F	4	70	45
G	9	95	25
H	2	75	50
I	9	100	20
J	6	80	35

(a) Measure the correlation of tensile strength and number of strands as indicated by this sample.

(b) Measure the correlation of tensile strength and weight as indicated by this sample.

(c) Chart each of the regressions involved in these measures.

(d) Calculate the standard error of estimate for each of these relationships.

(e) Calculate a coefficient of multiple correlation showing the relationship of tensile strength to both the number of strands and the weight of the cord.

(f) Estimate the probable strength of a sample having 83 strands and weighing 37 ounces.

7. The problem faced by the management is to discover conditions that are associated with voluntary quits and, on the basis of such analysis, to predict probable continuance rates and necessary replacements. To these ends, quit rates are compared with actual working hours per week and with compensated overtime. The following data are a sample of that with which analysis proceeds:

Period	Quit rates (Y)	Weekly hours (X_1)	Overtime (X_2)
a	0.5	21	2
b	0.6	24	4
c	0.7	22	1
d	0.8	30	5
e	0.9	26	7
f	1.0	32	6
g	1.1	40	5
h	1.2	35	8
i	1.5	36	7
j	1.7	34	5

(a) Calculate the simple correlation coefficients, r_{yx_1}, and r_{yx_2}.

(b) Calculate the coefficient of multiple correlation, $R_{yx_1x_2}$.

(c) Making use of the multiple regression, estimate values for quit rates and compute errors of estimate.

8. The following data compare wage rates in several industries with accident frequency and severity rates in the same industries. It is desired to discover to what extent wages compensate for hazards involved.

(a) Measure the correlation of wages and frequency rates.

(b) Measure the correlation of wages and severity rates.

(c) Secure a measure of multiple correlation involving both frequency and severity rates and wages.

Industry	Y Hourly wages	X_1 Frequency rates	X_2 Severity rates
Automobiles	46.5	19.41	1.02
Clay products	24.7	27.10	1.33
Cement	29.5	4.79	2.39
Leather	31.6	13.66	0.43
Lumber	20.8	59.67	5.00
Paper and pulp	32.6	19.47	1.70
Petroleum	40.7	12.85	1.89
Meat packing	32.3	30.81	1.19

9. Given the following sample of data respecting the employees of the G company:

(a) (1) Times tardy	(b) (2) Years of service	(c) (0) Weekly wages
26	10	25
15	6	11
16	9	16
18	10	33
26	12	36
27	13	37
17	11	30
15	9	12

(a) Calculate r_{ab}. How would you interpret this relationship?

(b) Calculate r_{ac}. How would you interpret this relationship?

(c) Calculate r_{bc}. How would you interpret this relationship?

(d) Calculate the partial regression coefficients or weights (b_1 and b_2), assuming that c is dependent, and evaluate their significance.

(e) Calculate $R_{c.ab}$.   How may its significance be described?

10. As a part of its training program, the C Company maintains an arrangement whereby its employees in certain departments are encouraged to undertake regular course work at the University of ——.   In an effort to evaluate this feature of training, the personnel division prepared the following tabulation comparing the hours of credit so earned by various employees with their production records and their scores on an intelligence test administered by the division:

Employee	Y Production: Average weekly wage	X_1 Hours of credit	X_2 Test scores
1	$ 9.00	9	90
2	11.00	6	95
3	12.50	3	99
4	14.00	3	95
5	15.00	12	92
6	17.50	9	108
7	20.00	12	103
8	22.50	15	91
9	24.00	6	98
10	25.00	3	94
11	25.00	12	97
12	25.00	9	89
13	27.50	9	86
14	28.00	15	93
15	28.00	12	96
16	29.50	15	86
17	30.00	12	88
18	30.00	21	85
19	30.00	24	87
20	32.50	15	79
21	32.50	21	95
22	34.00	27	74
23	35.00	15	85
24	35.00	18	73
25	36.00	18	82
26	37.50	15	83
27	38.00	18	70
28	40.00	15	72
29	42.50	30	87
30	45.00	27	84

(a) Ignoring other factors to be taken into account in further analysis of these data, evaluate the apparent relationship between the amount of credit earned under the circumstances described and production as measured by weekly earnings and express this relationship in terms of a coefficient of correlation (r).

(b) In a similar manner evaluate the relationship between test scores and production.

(c) Discover the multiple coefficient of wages on credits and test scores, and the prediction equation.

11. The Y Company seeks to perfect a means of identifying young men likely to become successful bond salesmen, so that its expensive training program may be made available only to the most promising candidates. It has measured the relationship between sales volume in the first year of active solicitation (which it considers a satisfactory measure of success) and a number of possible predictive conditions. The measures thus derived may be summarized as follows:

Linear correlation of sales (in thousands of dollars) with:

1. College grades; $r_{z_1 y} = 0.80$.
2. Entrance-test scores; $r_{z_2 y} = 0.90$.
3. Years of business experience; $r_{z_3 y} = 0.15$.
4. Reference ratings; $r_{x_4 y} = 0.10$.
5. High-school grades; $r_{z_5 y} = 0.21$.

The management concludes that only the first two of these measures are sufficiently important to be included in further analysis. It considers, therefore, the following additional data with respect to these conditions:

$\sigma_y = 0.08$.	$N = 100$.	$M_y = 5$.
$\sigma_{x_1} = 0.5$.	$r_{x_1 x_2} = 0.60$.	$M_{z_1} = 70$.
$\sigma_{x_2} = 1.0$.		$M_{z_2} = 80$.

(a) Combine the two simple correlations and measure the association between sales and the two independent series, college grades and entrance-test scores.

(b) Predict probable sales for an individual whose average college grade is 80 and whose entrance-test score is 80.

(c) Find k and σ_d, and indicate the limits within which such sales may be expected, with practical certainty, to fall.

12. Given the following measures of correlation, derived by preliminary statistical analysis and indicating the covariation of sales (series 0) and certain other conditions in 100 individual retail stores:

Simple coefficient of correlation of sales with the number of salesmen employed, $r_{01} = 0.72$.

Simple coefficient of correlation of sales with the amount of floor space, $r_{02} = 0.65$.

Intercorrelation of number of salesmen and amount of floor space, $r_{12} = 0.60$.

(a) Calculate the multiple correlation coefficient $R_{0.12}$.

(b) Calculate the partial coefficient $r_{01.2}$.

(c) Calculate the partial coefficient $r_{02.1}$.

13. Given the following simple coefficients of correlation, for three series:

$$r_{01} = 0.80$$
$$r_{02} = -0.40$$
$$r_{12} = -0.10$$

(a) Calculate the multiple coefficient, $R_{0.12}$.

(b) Calculate the partial coefficient $r_{01.2}$.

(c) Calculate the partial coefficient, $r_{02.1}$.

14. A small specialty manufacturing firm is seeking to forecast the demand for its products in communities in which they have yet to be introduced by means of the calculations shown below. It has compared the quantities of the product sold in each of 25 counties with (1) mail inquiries received from that area in the first month of the sales campaign and (2) the number of radio sets owned in the same county. The data thus collected appear as follows:

County	Number of items sold	Radios owned	Mail inquiries
	X_0	X_2	X_1
A	1,500	200	5,000
B	2,500	1,200	5,500
C	1,700	500	5,700
D	1,800	300	5,800
E	3,000	700	6,000
F	1,600	400	6,200
G	1,900	500	6,500
H	2,000	800	6,600
I	4,000	1,800	6,700
J	2,100	300	6,900
K	2,400	100	7,000
L	2,600	300	7,200
M	3,000	400	7,400
N	5,000	900	7,600
O	3,200	800	7,800
P	3,400	600	8,000
Q	3,600	500	8,200
R	3,200	500	8,400
S	6,000	300	8,600
T	4,200	1,600	8,800
U	2,700	1,300	9,000
V	4,500	1,100	9,200
W	4,900	300	9,300
X	4,200	1,500	9,600
Y	5,000	1,400	9,700

(*a*) Calculate the coefficient of correlation that measures covariation in sales and mail inquiries (r_{01}).

(*b*) Plot the regression of sales on inquiries, and indicate the standard error of estimate for this regression.

(*c*) Calculate the coefficient of correlation that measures covariation in sales and in owned radios.

(*d*) Plot the regression of sales on radios, and indicate the standard error of estimate for this regression.

(*e*) Calculate the coefficient of multiple correlation.

(*f*) Compare the standard error of estimate for the multiple regression with the other standard errors of estimate previously noted.

(*g*) Making use of the multiple regression equation, estimate probable sales for a county from which 8,500 mail inquiries are received and in which there are 650 radios.

(*h*) What are the limits within which such an estimate may be made with practical certainty (3 standard errors of estimate)?

15. Douglas and Cobb ("A Theory of Production," *American Economic Review*, Supplement, March, 1928, pages 139–165) calculated index numbers of deflated value of capital used in manufacturing ($C = X_1$), the number of workers employed in manufacturing ($L = X_2$), and an index of the physical product of manufactures ($P = X_0$) in the United States for the years 1899–1922, as listed below. A chart of these data will indicate that the relationship is logarithmic. Using the logarithms of the data, compute a multiple correlation of capital (C) and labor (L) as independent series with manufacturing product (P) as the dependent series ($R = 0.958$).

If it is assumed that b_1 and b_2, the coefficients applicable to the logarithms of capital and labor respectively, add to unity, the regression equation of multiple correlation reduces to

$$\log T = \log a + b \log C + (1 - b) \log L$$

and the normal equations become

$$N \log a + b\Sigma(\log C - \log L) = \Sigma(\log P - \log L)$$

$$\log a \Sigma(\log C - \log L) + b\Sigma(\log C - \log L)^2 = \Sigma(\log P - \log L)(\log C - \log L)$$

where b is the coefficient of the capital series ($\log X_1$) and $1 - b$ is the coefficient of the labor series ($\log X_2$). The regression equation thus obtained may be reduced from the logarithmic form by restating it as

$$T = aC^b L^{1-b} = 1.01 C^{0.25} L^{0.75}$$

INDEXES OF MANUFACTURING, UNITED STATES

(Adapted from Douglas and Cobb, *op. cit.*, by permission.)

Year	Capital		Labor		Product	
	Index	Log	Index	Log	Index	Log
1899	100	.0000	100	.0000	100	.0000
1900-	107	.0294	105	.0212	101	.0043
1901	114	.0569	110	.0414	112	.0492
1902	122	.0864	118	.0719	122	.0864
1903	131	.1173	123	.0899	124	.0934
1904	138	.1399	116	.0645	122	.0864
1905	149	.1732	125	.0969	143	.1553
1906	163	.2122	133	.1239	152	.1818
1907	176	.2455	138	.1399	151	.1790
1908	185	.2672	121	.0828	126	.1004
1909	198	.2967	140	.1461	155	.1903
1910	208	.3181	144	.1584	159	.2014
1911	216	.3345	145	.1614	153	.1847
1912	226	.3541	152	.1818	177	.2480
1913	236	.3729	154	.1875	184	.2648
1914	244	.3874	149	.1732	169	.2279
1915	266	.4249	154	.1875	189	.2765
1916	298	.4742	182	.2601	225	.3522
1917	335	.5250	196	.2923	227	.3560
1918	366	.5635	200	.3010	223	.3483
1919	387	.5877	193	.2856	218	.3385
1920	407	.6096	193	.2856	231	.3636
1921	417	.6201	147	.1673	179	.2529
1922	431	.6345	161	.2068	240	.3802

CHAPTER XVII

CURVILINEAR CORRELATION

All the measures of covariation of simple, multiple, and partial correlation described in preceding chapters are founded upon the assumption that the association between the related variables is linear, i.e., that the relationship may be described by a straight regression line. Such a regression implies that the variables are simple functions of each other and, as such, increase or decrease by regular increments (depending upon whether the correlation is positive or negative).

The linear relationship thus described is entirely appropriate for many types of association, as has been indicated in the chapter on simple correlation. In physics, for example, the relationship between successive units of time and the total distance traveled by steadily moving vehicles is accurately pictured by a straight-line trend. Similarly, the relationship between units of production and wages based upon a straight piece rate is of the same nature.

Curvilinear regressions.—In many instances, straight regression lines do not realistically represent the type of covariation involved. For example, the total distance traversed by a falling body in successive periods of time is described by a parabola rather than by a straight line, because of the factor of acceleration. In a similar manner, there are many instances of non-linear covariation in the field of business statistics. Practically all fluctuations in the demand for and supplies of various commodities that accompany variations in price, for instance, require curvilinear representation, as does the relationship between such conditions as rainfall and crop yields, mentioned in a preceding chapter. Similar, also, is the relationship of interest rates and the volume of loans, and that of increases in production accompanying increases in capital used in produc-

tion.   To illustrate this situation again, it may be indicated that, if daily production per worker is compared for various lengths of workday, such production may be expected to expand as hours increase up to what is regarded as usual or normal, but expansion will probably not continue at the same rate if hours are extended beyond these limits.   The regression for the whole association must, therefore, be represented by a curve.   The relationship between wages and production is also curvilinear rather than linear when workers are paid on the basis of an incentive wage, in which the piece rate is varied as production increases.

In these cases, it is clear that no simple linear relationship can accurately describe the association between variables, and it is equally clear that any correlation coefficient based upon a simple linear measure of association undervalues the actual covariation in the series, since a smaller proportion of the total variation is then accounted for than when a more representative regression line is used.   It is desirable, therefore, that the measure of association and the regression upon which it is based be adapted, in such cases, to the nature of the relationship.   The process by means of which such adaptation is effected is known as *curvilinear correlation.*

Curvilinear and linear correlation compared.—Curvilinear correlation is not essentially different from linear correlation. In both, the basic process consists of (1) discovering some pattern of covariation that expresses the regression, (2) calculating the amount of variation in the dependent variable that is accounted for by this pattern, and (3) comparing this "accounted-for" variation with the total to be explained.   In ordinary linear correlation, this result is achieved (although the basic process may be obscured by the methods of short-cut correlation procedure) by fitting a least-squares straight-line regression to the data and estimating values on the basis of this pattern of covariation.   The variance in these estimated or "accounted-for" values is then compared with the actual variance to be accounted for, and the ratio of the former to the latter is described as the *coefficient of determination.*

In curvilinear correlation, the pattern of association is por-

trayed by an appropriate regression curve. The particular curve to be selected depends upon the nature of the data, and the best guide to its selection is an understanding of the actual relationships between the series whose association is to be measured. In many cases, no mathematical statement of such a relationship is known, and the kind of curve is dictated by the nature of the data as they appear in a scatter diagram.

Because inspection must be generally relied upon as a basis for the selection of a regression curve, a few essential principles of conservative statistical practice must be borne in mind. In the first place, it is desirable to chart the data, as a preliminary to any sort of extensive or intensive statistical analysis. Such a graphic portrayal will frequently answer some of the more simple questions of interrelationships directly, and it may offer valuable suggestions as to what types of statistical analysis may prove worth while. In the second place, good judgment plays a large part in the selection of an appropriate curve upon which to base measures of correlation. Obviously, if sufficiently complex curves were used, they might be made to pass through every point in a given scatter diagram, but the resulting measure of correlation would be entirely meaningless. No amount of manipulation of data, mathematical or otherwise, can, therefore, take the place of sound common sense.

Selection of the regression curve.—The first step in curvilinear correlation is the selection of an appropriate regression curve. The problem involved, unless the nature of possible relationships between the variables is known or strongly suspected, is identical with that of fitting a trend, a problem given extensive consideration in Chapters X and XI. The curve may be fitted freehand, and an extensive correlation procedure, highly valuable even though it represents an approximation rather than a mathematically determined measure of association, has grown up about freehand curves used as regressions.[1] Some of the elementary principles of such analysis are given attention in a later portion of this chapter.

---

[1] This methodology has been developed most effectively by the agricultural economists. See, for instance, L. H. Bean, "Graphic Curvilinear Correlation," *Journal of the American Statistical Association*, 24, December, 1929, pp. 386–398.

Frequently, however, knowledge of the nature of the data under consideration and an understanding of the types of curves represented by various mathematical formulas suggest, as appropriate, one of the more common types of mathematically definable curves pictured in Fig. 10·3 (see page) 227. Thus, it may be that a second-degree parabola or an exponential is the curve of best fit. Sometimes, a straight line or a relatively simple curve may be satisfactory when one of the scales is adjusted to represent logarithms or reciprocals of the original data. What is essential, as far as this step in the process is concerned, is that the selection of the curve be recognized for what it is, i.e., primarily a matter of judgment, excepting in those rare cases where the exact nature of the data is known and forms a basis for the mathematical statement of interrelationship.

This first step in curvilinear correlation procedure may be illustrated by reference to the same data as that used in the preliminary discussion of correlation analysis, data representing a comparison of sales with psychological test scores (see Table 14·1, page 327). Data have been selected to represent fairly uniform covariation throughout the range, so that a linear regression is not unsatisfactory, but it will appear from inspection of the charted data that there is a slight tendency for sales to increase less rapidly in the higher ranges than in the lower ranges. Under these circumstances, a regression curve, parabolic in shape, represents a slightly better portrayal of the association than the straight line fitted in Chapter XIV. Hence, the measure of covariation afforded by such a regression line should be somewhat higher than that secured as the linear coefficient of correlation.

Fitting the parabolic regression.—The fitting of the parabolic regression curve by the method of least squares is fundamentally the same as the trend-fitting process described in Chapter XI. As has been suggested, the elements in this method are designated as X_1 and X_2, because, in the process of centering each series at its mean, the relationship between the elements is broken down. Hence, usage here follows precisely the procedure of multiple correlation as described in Example 16·3, page 383. The two independent elements are designated as

X_1 and X_2. The normal equations as applied to these data when they are expressed as deviations from the mean of each series reduce to the same form as that employed in multiple correlation, namely

$$b_1 \Sigma x_1^2 \quad + b_2 \Sigma x_1 x_2 = \Sigma x_1 y$$
$$b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 \quad = \Sigma x_2 y$$

It may be desirable to substitute b for b_1 and c for b_2 in order

EXAMPLE 17·1

FITTING THE PARABOLIC REGRESSION CURVE

Data: See Table 14·1, page 327.

Block	Row / Salesmen	(1) X_1	(2) $X_2 (= X_1^2)$	(0) $X_0 (= Y)$	Z
	A	4	16	5	25
	B	5	25	4	34
	C	6	36	5	47
	D	4	16	6	26
	E	5	25	9	39
	F	6	36	10	52
	G	6	36	9	51
	H	7	49	12	68
	I	9	81	11	101
	J	8	64	9	81
	S	60	384	80	524
P	1	384	2,610	508	3,502
	2		18,708	3,420	24,738
	0			710	4,638
	Z		·		$32,878 = \Sigma Z^2$
Np	1	240	3,060	280	3,580
	2		39,624	3,480	46,164
	0			700	4,460
	Z				$54,204 = N \Sigma z^2$

By solution of the normal equations, the trend is defined as (see Example 17·2):

$$T = -4.63047 + 3.05087 X_1 - 0.14778 X_1^2$$

to suggest the parabolic form, and to utilize X_0 for the dependent element, Y, in which case the equations take the form

$$b\Sigma x_1^2 \quad + c\Sigma x_1 x_2 = \Sigma x_1 x_0$$
$$b\Sigma x_1 x_2 + c\Sigma x_2^2 \quad = \Sigma x_2 x_0$$

In Examples $17\cdot1$ and $17\cdot2$, the fitting of the parabolic regression curve is accomplished in a somewhat condensed form. Test scores are represented by X_1, and the squares of these scores are represented by X_2. The sales records appear as X_0. Below the columns of data are the summed products in block P. For example, the item *384* in row 1, column 1, of this block is the sum of the squares of items in this column. Similarly, the item *2610* in row 1, column 2, is the sum of the products of columns 1 and 2, etc.

Directly below the summed products in block P are the centered summed products multiplied by N, which are obtained by the usual centering formulas of the type:

$$N\Sigma x_1^2 \quad = N\Sigma X_1^2 \quad - (\Sigma X_1)^2$$
$$N\Sigma x_1 x_2 = N\Sigma X_1 X_2 - \Sigma X_1 \Sigma X_2$$

Final solution, as is indicated in the example, may utilize the Doolittle method, or it may involve substitution in the normal equations, or it may be effected by means of formulas [1]

[1] These formulas, derived from the centered normal equations, are

$$c = \frac{\Sigma x_1^2 \Sigma x_2 x_0 - \Sigma x_1 x_2 \Sigma x_1 x_0}{\Sigma x_1^2 \Sigma x_2^2 - \Sigma x_1 x_2^2} = \frac{240 \times 3{,}480 - 3{,}060 \times 280}{240 \times 39{,}624 - 3{,}060^2} = -0.14778$$

$$b = \frac{\Sigma x_1 x_0 - c\Sigma x_1 x_2}{\Sigma x_1^2} = \frac{280 - (-0.14778 \times 3{,}060)}{240} = 3.05086$$

If the original normal equations are used (cf. page 375), they become

$$10a + \quad 60b + \quad 384c = \quad 80$$
$$60a + \quad 384b + \quad 2{,}610c = \quad 508$$
$$384a + 2{,}610b + 18{,}708c = 3{,}420$$

The values thus determined are: $a = -4.63047$; $b = 3.05087$; $c = -0.14778$.

If the centered figures, times N, are used, the abbreviated equations (page 408) become

$$240b + \quad 3{,}060c = \quad 280$$
$$3{,}060b + 39{,}624c = 3{,}480$$

which yield: $b = 3.05087$; $c = -0.14778$, and a is found by the first normal equation, as

$$Na + b\Sigma X_1 + c\Sigma X_2 = \Sigma Y$$
$$10a + 3.05087 \times 60 - 0.14778 \times 384 = 80$$
$$a = -4.63047$$

<div align="center">

EXAMPLE 17·2

SOLUTION OF NORMAL EQUATIONS BY DOOLITTLE METHOD[1]

(The summations are multiplied by N; see Example 17·1)

</div>

Directions:	Block	Row	(1)	(2)	(0)	Z	
From Ex. 17·1	Np	1	240	3,060	280	3,580	
		2		39,624	3,480	46,164	
		0			700	4,460	
Enter first Np	I	s_1	240	3,060	280	3,580	
$-s_1/(s_1$ col. 1)		q_1	−1	− 12.75	−1.16667	−14.91667	Check
Enter second Np	II	1		39,624	3,480	46,164	
s_1 (q_1 col. 2)		2		−39,015	−3,570	−45,645	
Add—		s_2		609	−90	519	Check
$-s_2/(s_2$ col. 2)		q_2		−1	0.14778	−0.85222	Check
Enter q_1	C	q_1	−1	− 12.75	−1.16667		
Enter q_2		q_2		−1	0.14778		
Enter −1 and b's		b	3.05087	−0.14778	−1		
b times q_1		bq_1		1.88420	1.16667	Sum = b_1	
b times q_2		bq_2			−0.14778	= b_2	
Enter S (Ex. 17·1)	A	S	60	384	80		
b times S		bS	183.05220	−56.74752	−80		
$-Na + (-N)$		a				46.30468 = $-Na$	
						−4.6305 = a	
Enter Col. 0 of Np	R	1	$N\Sigma x_1 x_0 = 280$	$N\Sigma x_2 x_0 = 3,480$	$N\Sigma x_0^2 = 700$		
$(b_1\Sigma x_1 x_0 + b_2 \Sigma x_2 x_0)$		2	$(854.24360 - 514.27440) + 700 = 0.48567 = \rho^2$				
$+ \Sigma x_0^2$		3	$\rho = \sqrt{0.48567} = 0.697$				

NOTE: It will be seen that the last two blocks are solutions of the equations

$$Na = \Sigma Y - b_1 \Sigma X_1 - b_2 \Sigma X_2$$
$$\rho^2 = (b_1 \Sigma x_1 x_0 + b_2 \Sigma x_2 x_0) \div \Sigma x_0^2$$

[1] Most of the abbreviations will be self-evident. In block I the first row of Np is entered and labeled s_1. It is then divided by 240 taken negatively, or, in symbols, $-s_1/(s_1$ col. 1). In block II the second Np row is entered as it stands, i.e., *not* taken as a "full row," as it was in determining the row total, Z. The next line in block II is the row s_1 (beginning in the second column) times the second item of q_1, that is, q_1 col. 2, or −12.75. Row s_2 is the sum of the two preceding rows. Row q_2 is computed as is q_1. In block C the constants are calculated from rows q_1 and q_2, which are entered for convenience. The computation begins with entering, in row b, −1. This, times the two items above it, gives the items below it. The item thus obtained in row bq_2 is b_2, which is entered in row b, column 2. Row bq_1 can then be completed by the product $(-12.75)(-0.14778) = 1.88420$. Its sum, b_1, is entered in row b, column 1. The computation of a and ρ can be made as part of the Doolittle method, as in blocks A and R, or the usual equations, indicated below block R, may be directly employed.

prepared for that purpose. It will be noted that the constants b_1 and b_2 are here referred to as b and c to conform to earlier usage in the fitting of the parabola. The value of a is found as usual by means of the first normal equation, which may be used in either of the forms:

$$Na = \Sigma Y - b\Sigma X - c\Sigma X^2$$
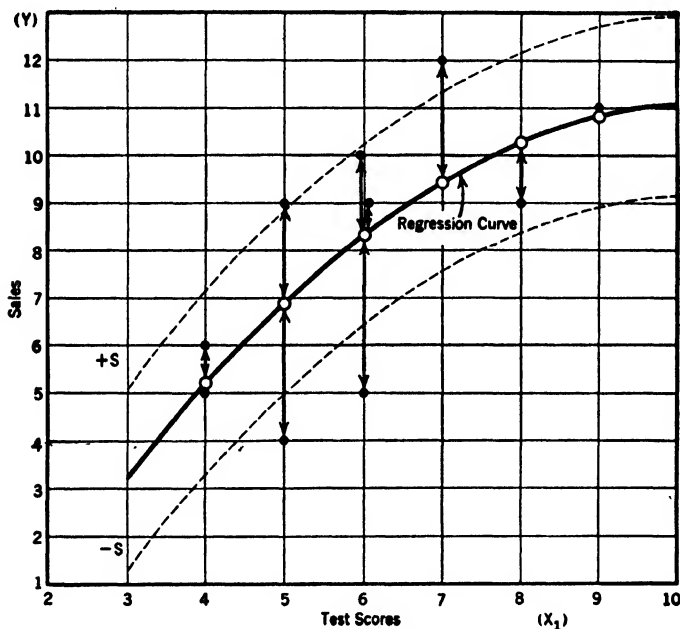
or

$$a = M_0 - bM_1 - cM_2$$

FIG. 17·1.—Parabolic Regression of Sales Records on Test Scores (see Examples 17·1 and 17·2).

where M_2 is the mean of X_1^2. This equation may be included in the Doolittle solution, as is indicated in Example 17·2.

The regression equation is found to be

$$T = -4.6305 + 3.0509X_1 - 0.1478X_1^2$$

and the trend it describes is graphically represented in Fig. 17·1.

After the regression equation has been determined, estimates of sales based on the curvilinear relationship to test scores may

be made by the usual method of substituting the given values of X in the equation. The results of such substitutions make up the series of points indicated on the regression line in the figure already mentioned. The curvilinear regression may be compared with the earlier linear trend fitted to the same data (see pages 328 and 340). It will be noted that, although the curvature of the new regression is not great, it effects a slightly better fit than the linear trend.

Measurement of curvilinear correlation.—As is true of all the other types of correlation described in preceding pages, the regression thus defined is regarded as "accounting for" a portion of the variance in the dependent variable, sales, and the measure of correlation is the square root of the ratio of this "accounted-for" variance to the total variance to be explained. In curvilinear correlation, however, the ratio σ_t/σ_0 is called the *index of correlation*, in order to distinguish it from the linear *coefficient*, and its symbol is the Greek rho, ρ.

The coefficient might be calculated directly by means of the T values, that is, the estimates of Y. The standard deviation of these estimates may be calculated, and divided by the standard deviation of the dependent series (X_0 or Y) for the decimal fraction that represents the index. This process, however, is tedious and subject to error, and in actual practice various short cuts are found convenient. Possibly the simplest of these is that employing the formula

$$\Sigma t^2 = b\Sigma x_1 x_0 + c\Sigma x_2 x_0$$

In the illustrative example, Σt^2, so calculated, is

$$\Sigma t^2 = (3.05087)(28) + (-0.14778)(348) = 33.99692$$

The square of the index of correlation is then found as the ratio of Σt^2 to Σx_0^2 (in Example 17·1, $N\Sigma x_0^2$ was found to be 700).

$$\rho^2 = \frac{\Sigma t^2}{\Sigma x_0^2} = \frac{33.99692}{70} = 0.486$$

and

$$\rho = \sqrt{0.486} = 0.697$$

The indexes of determination, non-determination, and aliena-

tion bear the same relationship to this index of correlation as the coefficients of the same order bear to the coefficient of correlation, and they have the same significance and meaning.[1]

Statistical reliability of the index of correlation.—The statistical reliability of the index of correlation may be broadly estimated by reference to its standard error, calculated in the usual manner as

$$\sigma_\rho = \frac{(1 - \rho^2)}{\sqrt{N}}$$

or, if allowance is to be made for the error of sampling, the divisor may be $\sqrt{N - m}$. The meaning of this standard error is the same as that attached to the standard errors of simple and multiple coefficients of correlation. The standard error of the index measured in preceding paragraphs is

$$\sigma_\rho = \frac{(1 - \rho^2)}{\sqrt{N - m}} = \frac{1 - 0.48567}{\sqrt{10 - 3}}$$

$$= \frac{0.51433}{\sqrt{7}} = 0.19440.$$

The breadth of this range indicates the limitations imposed by the number of observations and the lack of confidence that can be placed in the index. However, this standard error of the index is not satisfactory unless N is reasonably large (at least 100 items).

As has already been explained, a superior method of evaluating the reliability of measures of correlation is that provided by tables or charts indicating the chance measures that would occur under specified conditions. According to the appropriate chart shown on page 559, it will be found that, for 10 items and

---

[1] If the centered squares have not been found, ρ^2 may be written as

$$\rho^2 = \frac{\Sigma t^2}{\Sigma x_0^2} = \frac{N\Sigma T^2 - (\Sigma X_0)^2}{N\Sigma X_0^2 - (\Sigma X_0)^2}$$

and ΣT^2 is found as

$$\Sigma T^2 = a\Sigma Y + b\Sigma XY + c\Sigma X^2 Y$$

or

$$\Sigma T^2 = a\Sigma X_0 + b_1\Sigma X_1 X_0 + b_2\Sigma X_2 X_0$$

3 variables,[1] an index greater than 0.76 could appear only once in 20 times purely by chance from uncorrelated data. This so-called 5 per cent limit is conventionally regarded as the lower limit of reliability. The index discovered in preceding paragraphs, $\rho = 0.697$, is substantially below this figure. Hence, the result obtained in this illustration by the process of curvilinear correlation does not measure up to acceptable statistical standards and cannot, therefore, be relied upon as a means of estimating or predicting sales performance from the data of psychological test scores.[2]

Curvilinear multiple correlation.—The curvilinear type of correlation analysis is applicable to multiple as well as simple covariation. To illustrate the procedure involved, reference may be made to the same data as those used in demonstrating the process of multiple linear correlation in the preceding chapter, but it should be remembered that these data have been simplified, for purposes of exposition, to the point where they cannot be expected to provide statistically significant results. The items represent, as the dependent variable, sales expressed in thousands of dollars, and, as independent variables, test scores and years of experience of the salesmen involved. They are summarized in tabular form on page 372.

The first problem is to fit a curvilinear multiple regression to these data, and the principle utilized in solving this problem is not different from that involved in fitting linear and curvilinear regressions. It is only necessary to set up a hypothesis involving more elements. If it is assumed, as before, that parabolic regressions will most adequately measure the curva-

---

[1] These variables are Y, X, and X^2, or they may be regarded simply as the number of constants in the regression equation.

[2] As in other types of correlation, a more general appraisal may be made by means of another measure of correlation, the statistic F. The 5 and 1 per cent levels of its sampling distribution are given for various degrees of freedom in the table on page 586. It is computed as

$$F = \frac{\rho^2}{1 - \rho^2} \times \frac{N - m}{m - 1} = \frac{0.4857}{0.5143} \times \frac{10 - 3}{3 - 1} = 3.33$$

where m is the number of constants (a, b, and c) in the regression equation. The least significant value of F in this case (row 7, col. 2) is 4.74. For a description of this measure, see Appendix, pages 554–556.

ture involved, and, if both the psychological-test scores and the years of experience are to be taken as independent series, then the elements of the regression become

$$1, \ X_1, \ X_2, \ X_1^2, \ \text{and} \ X_2^2$$

and the problem consists of finding suitable weights (constants) to be applied to these elements so that their weighted total will conform as closely as possible (according to the least-squares criterion) tö that of items of the independent series, sales, designated in this case X_0. The weight-finding process involves the usual dependence upon normal equations, although in practice, with such an extensive number of constants, it is seldom feasible actually to substitute in these equations to discover values. Rather, dependence must be placed upon a more automatic type of procedure, such as the Doolittle method.

The complete regression equation is

$$T = a + b_1 X_1 + b_2 X_2 + c_1 X_1^2 + c_2 X_2^2$$

But, for convenience in calculation, the last two terms should be designated as new independent series, making the equation

$$T = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$$

The values of the constants may be computed by means of normal equations which are merely expansions of those previously used. The first is the trend equation, summed and equated to ΣY. The second is the first multiplied through by X_1, thus:

(1) $\qquad na + b_1 \Sigma X_1 + b_2 \Sigma X_2 + b_3 \Sigma X_3 + b_4 \Sigma X_4 = \Sigma Y$

(2) $\qquad a \Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_1 X_2 + b_3 \Sigma X_1 X_3 + b_4 \Sigma X_1 X_4$
$$= \Sigma X_1 Y$$

and the remaining three are the first multiplied by X_2, X_3, and X_4, respectively. The solution may involve direct substitution in these normal equations,[1] or they may be centered

---

[1] If the normal equations are directly employed, P may be found as

$$P = \frac{\Sigma t^2}{\Sigma x_0^2} = \frac{N \Sigma T^2 - (\Sigma X_0)^2}{N \Sigma X_0^2 - (\Sigma X_0)^2}$$

where ΣT^2 is found as

$$\Sigma T^2 = a \Sigma X_0 + b_1 \Sigma X_1 X_0 + b_2 \Sigma X_2 X_0 + b_3 \Sigma X_3 X_0 + b_4 \Sigma X_4 X_0$$

as were those employed in multiple correlation. The solution is preferably carried out by the Doolittle method, as is illustrated in the Appendix, pages 540–541. It results in an index of multiple correlation (symbol, capital rho) and a value of F of

$$P = 0.870; \quad F = 3.89$$

Measures of reliability follow the same procedure as has already been explained in connection with Example 17·1, page 408. It will be seen (pages 559 and 586) that the least significant P is 0.898, and the least significant F is 5.19.

The description of the measurement of curvilinear correlation in preceding paragraphs has been sufficiently extensive to indicate that it increases the amount of calculation involved and requires a more adequate basis in the number of observations than does the measurement of linear relationships. There are many situations, however, in which no accurate or satisfactory measure of covariation can be secured unless the curvilinear measure is applied. In some cases, curvilinear measures may be made for one or two independent series, and others may be satisfactorily represented by linear regressions. It may not be possible to tell at a glance which series require curved regressions, so that both linear and curvilinear measures may have to be made, unless knowledge of the nature of the data itself is conclusive on this point.

Curvilinear correlation of grouped data.—Curvilinear correlation of two series of data that have been tabulated in a correlation table or bivariate scatter diagram may be effected by a process similar to that employed in the simple linear correlation of tabulated data. When data to be analyzed are numerous, it is often convenient to incorporate them in such a correlation table, although the correlation procedure is somewhat complicated by the presence of the frequencies.[1] The process is illustrated in Example 17·3, where the index of correlation has been calculated for the tabulated and coded data of Table 17·1.

---

[1] When extensive multiple or curvilinear correlations are to be found, it is advisable to secure the services of a statistical laboratory with adequate tabulating equipment.

TABLE 17·1

CORRELATION TABLE: CODED BIVARIATE DISTRIBUTION

Data: Assumed test scores and sales of a group of 20 employees (cf. Table 15·1, page 353) m_X and m_Y coded as X and Y.

Sales	Test Scores						
	m_X	3	5	7	9	11	
m_Y	Y \ X	0	1	2	3	4	f_Y
35	4				1	1	2
30	3			2	1	1	4
25	2		1	3	2		6
20	1		4	2			6
15	0	1	1				2
	f_X	1	6	7	4	2	20

After the data have been entered in their proper cells in the correlation table, the results of this tabulation are preferably set up in columns which represent the elements X_1, X_1^2, and X_0, as shown in Example 17·3. Thus the whole first row represents the item in the correlation table in the upper right-hand cell that has the value of 4 on the X scale and 4 on the Y scale; the second row of the table in Example 17·3 represents the other item in the first row of the correlation table, an item having an X or X_1 value of 3 and a Y or X_0 value of 4. It will be clear that the data thus taken from the correlation table have values that are expressed in terms of deviations from arbitrary origins in each scale, i.e., they are "coded" data, a fact that will

EXAMPLE 17·3

CURVILINEAR CORRELATION OF GROUPED DATA

Data: Assumed test scores and sales from Table 17·1.

f	X_1	X_2 or X_1^2	Y or X_0	Z
1	4	16	4	24
1	3	9	4	16
1	4	16	3	23
1	3	9	3	15
2	2	4	3	9
2	3	9	2	14
3	2	4	2	8
1	1	1	2	4
2	2	4	1	7
4	1	1	1	3
1	1	1	0	2
1	0	0	0	0
20	40	102	38	180

Block	Row	(1)	(2)	(0)	(Z)	
P	1	102	298	95	495	
	2		954	273	1525	
	0			98	466	$2486 = \Sigma Z^2$
Np	1	440	1880	380	2700	
	2		8676	1584	12140	
	0			516	2480	$17320 = N \Sigma z^2$

Centered normal equations with substitutions:

$$440b_1 + 1{,}880b_2 = \quad 380$$
$$1{,}880b_1 + 8{,}676b_2 = 1{,}584$$

Solving algebraically:

$$b_1 = 1.126904; \quad b_2 = -0.061616$$

Solving first normal equation for a:.

$$Na = \Sigma X_0 - b_1 \Sigma X_1 - b_2 \Sigma X_2$$
$$= 38 - 45.0762 + 6.2848 = -0.7914$$
$$a = -0.7914 \div 20 = -0.0396$$

Solving for ρ:

$$\rho^2 = (b_1 \Sigma x_1 x_0 + b_2 \Sigma x_2 x_0) + \Sigma x_0^2$$
$$= (428.224 - 97.600) + 516 = 0.6407$$
$$\rho = \sqrt{0.6407} = 0.800$$

not affect the measure of covariation but must be taken into account if the regression is to be used in prediction.

When the data have been set up in columnar form, the summed products are calculated as usual, but each summation takes into account the frequencies of the items. After these summed products are found, they are centered, i.e., expressed in terms of deviations from the means of each series, and the centered values (conveniently multiplied by N) are available for substitution in the centered normal equations. The solution of these equations may be effected in any one of the several ways, but the algebraic solution is probably the most convenient here, and it is the method utilized in the example. The values of a and ρ are secured by the usual formulas.

Decoding for prediction.—The regression equation for the data of Example $17 \cdot 3$ is found to be

$$T = -0.0396 + 1.1269X - 0.0616X^2$$

If, however, it is desired to use this equation in predicting sales from given values of the independent series, test scores, it will be necessary to "decode" the regression, since X and Y or X_1 and X_0 values are expressed in terms of arbitrary origins and class intervals. For instance, it will be seen by reference to Table $17 \cdot 1$ that, when coded Y is equal to zero, actual Y is 15 (i.e., $R_y = 15$) while the class interval in this series is 5 (i.e., $i_y = 5$). Thus the actual value of any Y item, or $\overline{Y}$, is 5 times its coded value plus 15, i.e.,

$$\overline{Y} = i_y Y + R_y$$

Hence, $\overline{T}$ or $\overline{Y}' = 5(-0.0396 + 1.1269X - 0.0616X^2) + 15$

$$= 14.8020 + 5.6345X - 0.3080X^2$$

But the X or X_1 values are similarly coded. It will be seen that when coded X is equal to zero, actual X, or $\overline{X}$, is 3 (i.e., $R_x = 3$) and the class interval in this series is 2 (i.e., $i_x = 2$).

It is obvious that

$$X = \frac{\overline{X} - R_x}{i_x}$$

and

$$\overline{T} = 14.8020 + 5.6345 \left(\frac{\overline{X} - 3}{2}\right) - 0.3080 \left(\frac{\overline{X} - 3}{2}\right)^2$$

$$= 5.6572 + 3.2792\overline{X} - 0.0770\overline{X}^2$$

In this form, the equation may be applied to given test scores to predict probable sales. If, for example, a test score of 6 is made by an applicant, his estimated sales will be measured by the equation
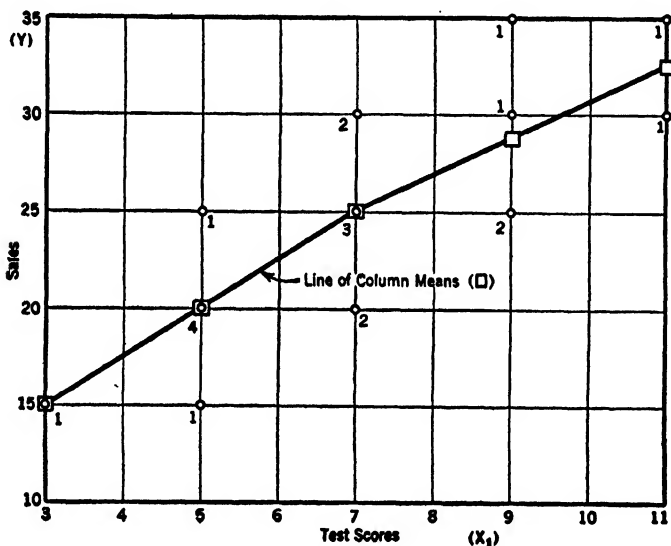
$$T = 5.6572 + 3.2792 \ (6) - 0.0770 \ (6)^2 = 22.5604$$

FIG. 17·2.—Regression of Correlation Ratio (Column Means) Fitted to Grouped Data (see Example 17·4).

Reliability of the regression.—The statistical reliability of the regression may be measured by reference to its standard error of estimate, calculated in the usual manner, and the reliability of the index of correlation may be evaluated by reference to its standard error, by consulting the chart of significance repeatedly used for this purpose, or by calculation of the statistic F. The chart indicates that an index of 0.55 is required to meet the 5 per cent standard, and an index of 0.65 is necessary

to meet the 1 per cent standard.   Hence, the index of 0.80, as calculated, may be regarded as reliable from the standpoint of sampling.

The correlation ratio.—It is frequently convenient to use a somewhat less formal method in measuring curvilinear covariation in tabulated data, a method which, while it does not provide an explicit regression equation, does furnish a measure of the degree of covariation.   It secures, as its measure of correlation, the *etā coefficient,* or *correlation ratio,* and it substitutes, for the usual mathematical regression curve, the averages of each of the columns of the scatter diagram or correlation table. (See Fig. 17·2.)   The correlation ratio is found as the grouped square root of the ratio of variance in column means to variance in the dependent series.   The scatter diagram is set up as usual, but preferably the classes are so arranged as to avoid small frequencies in the X distribution.

The correlation ratio should not be regarded as useful unless the number of items involved is large.   Coded scales will be found especially convenient, as is indicated in Example 17·4, where the data of Example 17·3 are again employed.   Two different approaches to the calculation of the eta coefficient are illustrated in parts A and B of Example 17·4.   The first is the more commonly described method; the second is a short cut that will be found extremely convenient in many cases, particularly since it provides an introduction to the analysis of variance, later to be discussed.

The first step of part A involves calculation of the average of the Y data and its standard deviation.   Procedure in this step is not different in any respect from that described in Chapter VI, except for the fact that coded values are used throughout in place of actual values of Y.   Then, the frequencies characteristic of each column are noted in the row "f_c," after which the totals of each column are calculated as ΣY_c.   Each column total is the sum of products of individual Y items and their frequencies. For example, the total of the second column is

$$(1)(1) + (4)(2) + (1)(3) = 12$$

The mean of each column is then found by dividing these totals

<div align="center">

EXAMPLE 17·4

THE CORRELATION RATIO OR ETA COEFFICIENT

</div>

Data: Sales (Y) and test scores (X) of Table 17·1, page 418.

A. Long method.

Class marks	Coded scale Y	3	5	X 7	9	11	f	d_y	fd_y	$f(d_y)^2$
				Frequencies						
35	5				1	1	2	2	4	8
30	4			2	1	1	4	1	4	4
25	3		1	3	2		6	0	0	0
20	2		4	2			6	-1	-6	6
15	1	1	1				2	-2	-4	8
f_c		1	6	7	4	2	20	20)-2	26	
ΣY_c		1	12	21	15	9		-0.1	1.3	
M_c		1	2	3	3.75	4.5		$R=3$		
$M_c-M_y=D$		-1.9	-0.9	0.1	0.85	1.6		2.9=M_Y		
D^2		3.61	0.81	0.01	0.7225	2.56				
f_cD^2		3.61	4.86	0.07	2.89	5.12	16.55 $= \Sigma f_cD^2$			

$$\sigma_{Mc} = \sqrt{\frac{\Sigma f_cD^2}{N}} = \sqrt{\frac{16.55}{20}} = \sqrt{0.8275} = 0.90967$$

$$\sigma_y = \sqrt{\frac{\Sigma f(d_y)^2}{N} - \left(\frac{\Sigma fd_y}{N}\right)^2} = \sqrt{\frac{26}{20} - \left(\frac{-2}{20}\right)^2} = \sqrt{1.29} = 1.13578$$

$$\eta_{yx} = \frac{\sigma_{Mc}}{\sigma_y} = \frac{0.90967}{1.13578} = 0.8009$$

B. Short method

Class marks	Coded scale Y	3	5	7	9	11	f	(ΣY) fY	(ΣY^2) fY^2
				Frequencies					
35	5				1	1	2	10	50
30	4			2	1	1	4	16	64
25	3		1	3	2		6	18	54
20	2		4	2			6	12	24
15	1	1	1				2	2	2
N_c or Σf_c		1	6	7	4	2	20	58	194 (Check)
ΣY_c		1	12	21	15	9	58 $(\Sigma Y)^2/N = 58^2/20 = 168.2$		
$\Sigma Y_c{}^2$		1	26	67	59	41	$194.00 - 168.20 = 25.80 = \Sigma y^2$		
$(\Sigma Y_c)^2/N_c$		1	24	63	56.25	40.5	$184.75 - 168.20 = 16.55 = \Sigma t^2$		
$\Sigma y_c{}^2$		0	2	4	2.75	0.5	$9.25 = \Sigma d^2;\ \eta = \sqrt{\Sigma t^2/\Sigma y^2}$		
							$= 0{:}80$		

by the frequencies characteristic of each column. The next step involves calculation of the differences between individual column means and the mean of the Y distribution, i.e., $M_c - M_y = D$, after which these differences are first squared and then multiplied by their appropriate frequencies, as indicated in successive rows in part A of the example. The total of the last-mentioned products is usually described as $\Sigma f_c D^2$, and it is used to calculate the standard deviation of the column means, which is σ_t, i.e.,

$$\sigma_t = \sqrt{\frac{\Sigma f_c D^2}{N}}$$

after which this standard deviation is compared with that of the Y series, as shown in the example, to secure the eta coefficient.

Since the correlation ratio, like other measures of correlation, is nothing more than a ratio of trend variability (trend is defined as the means of the columns) to Y variability, the short cut shown in part B is made possible through use of the usual centering equations, by means of which the manipulations are considerably simplified. The first steps involve calculation, by columns, of the variability, as indicated by the centered squares Σy_c^2. The sums of rows N_c, ΣY_c, and ΣY_c^2 obviously provide N, ΣY, and ΣY^2 for the whole table. They may be checked by utilizing the total Y frequencies, as indicated in columns to the right. The total of the correction terms $(\Sigma Y_c)^2/N_c$ is the sum of the squared regression points (ΣT^2). Both ΣY^2 and ΣT^2, since Y and T have the same sum, are centered by subtracting the correction term $(\Sigma Y)^2/N$. Eta squared is then found as

$$\eta_{yx}^2 = \frac{\Sigma t^2}{\Sigma y^2} = \frac{16.55}{25.80} = 0.6415$$

In practice the process may be abbreviated by omitting the third and fifth rows (ΣY_c^2 and Σy_c^2).

The coefficient thus defined, η_{yx}, represents the measure of covariation indicated by the regression of Y on X. It is also possible, of course, to measure η_{xy}, in which the regression of X on Y is involved. Frequently, in business statistics, the nature of the data will make the latter measure more or less meaning-

less, but the value may be readily secured by performing the manipulations on rows instead of columns.

Measures of unexplained variance, of σ_d corrected for sampling, and of the statistical reliability of the correlation ratio may be secured in the same way that similar measures have been obtained for coefficients and indexes of correlation.[1] If estimates are to be made, the average for each column should be regarded as the predicted value for any X item falling in that column. However, the method of fitting the regression line is such that little dependence should be placed on such estimates. As in all analysis involving tabulation, the choice of class limits may materially affect results, and great care must be exercised in this original step in the analysis.

Eta as a test of linearity.—The eta coefficient is always larger than the coefficient of linear correlation if the covariation being measured is curvilinear. Hence, the difference between the two measures may be used as an index of the linearity of covariation. This measure is sometimes calculated as "zeta" (ζ), where

$$\zeta = \eta^2 - r^2$$

If zeta approximates 0, the regression is linear; if zeta is materially greater than 0, the appropriate regression is presumably nonlinear. Thus, in the example used to demonstrate curvilinear correlation of grouped data, the fact that the eta coefficient of determination (η^2) does not vary greatly from the Pearsonian coefficient of determination (r^2) is indicative of fairly linear covariation.[2]

---

[1] In measuring the significance of η_{yx}, the number of columns is regarded as the number of variables, or constants in a potential regression equation passing through the mean of each column. In Example 17·4, therefore, where $N = 20$, the number of constants is 5. The value of eta that might be attained once in 100 times by mere chance is 0.75. (See Fig. A4, page 559.) For η_{xy}, the number of rows defines the number of variables.

[2] The significance of *zeta* may be measured by a method more fully discussed in the Appendix, involving the calculation of the statistic F by the formula (m is the number of columns):

$$F = \frac{\eta^2 - r^2}{1 - \eta^2} \times \frac{N - m}{m - 2} = \left(\frac{0.6415 - 0.6360}{0.3585}\right)\left(\frac{20 - 5}{5 - 2}\right) = 0.077$$

The fractional result indicates that *zeta* is smaller than would be expected even by

Coefficient of mean square contingency.—There are frequent occasions when it is desired to measure covariation between variables which allow classification into several categories but do not readily permit exact quantitative measurement. Pearson, in 1904, presented the method of contingency to meet the demands of such situations.[1] It is based upon the comparison of expected chance similarity between two variables and actual similarity, and makes use of the principle which states that, if two events are entirely independent, then the probability of their joint occurrence is the product of their separate probabilities. It is an adaptation of the measure of deviation called *chi square* (see Chapter XX), and is identical in principle with the measurement of fourfold correlation described in Chapter XV (cf. pages 361–363).

The usual measure of contingency is derived as a *coefficient of mean square contingency* (*CC*) which is based upon a tabular arrangement of the data similar to that employed in calculating η (see Example 17·4). There is an important difference, however, in that one or both of the scales may be qualitative rather than quantitative. For example, various degrees (very friendly, friendly, indifferent, hostile, or very hostile) of characteristic attitudes might be represented. It is obvious that such a scale cannot be accurately reduced to numerical terms, although an inspection of the frequencies might be made the basis of a hypothetical scale.[2] The calculation therefore does not involve X and Y scales, but merely the usual X and Y frequencies; that is, the frequencies by columns (f_c, one for each X class), and frequencies by rows (f_r, one for each Y class).

---

chance, and hence the regression is not significantly curvilinear. However, a result greater than 1 could be evaluated by reference to F table, page 586. The procedure is as follows: locate $N - m$, or 15, in the table stub, and select the column numbered $m - 2$, or 3. The values thus determined, 3.29 and 5.42, are the 5 per cent and 1 per cent levels of chance F, respectively.

[1] In the Appendix, page 549, two other convenient methods of measuring covariation in non-quantitative variables, the method of biserial r and that of biserial eta, are described.

[2] The $\Sigma f\%$ for each distribution could be plotted on probability paper, and L_2 set at such a spacing as to provide a straight line. The corresponding class marks would then be assigned to the frequencies.

The calculation of the coefficient is very simple. The first step is the squaring of each frequency and its division by the product of the appropriate column frequency and row frequency. These ratios are then added, and their sum, less 1, is designated as ϕ^2; that is,

$$\phi^2 = \Sigma\left(\frac{f^2}{f_c \times f_r}\right) - 1$$

The coefficient of mean square contingency is then found as

$$CC = \sqrt{\frac{\phi^2}{\phi^2 + 1}}$$

The usual procedure is illustrated in Example 17·5. The tabulation is set up in the usual form with the frequencies allocated to their proper cells. These are totaled by rows and columns to obtain f_r and f_c. The computation for each cell is given individually below the table, first as a fractional expression and, second, as a quotient. The sum of the ratios is 1.517. By means of the formula previously given, the coefficient is found to be 0.584.

Certain characteristics of this coefficient should be carefully noted. In the first place, it is very similar to the coefficient η in that it registers a high degree of correlation if the means of the columns are decidedly different or varying among themselves. It may be observed, however, that, while the upper limit of η and ρ is 1.00, the upper limit of CC merely approaches 1.00 as the number of classes is increased.[1] In a certain sense, therefore, it may be said that the coefficient, CC, penalizes a small number of classes by diminishing the upper limit.

The reliability of the coefficient of contingency is best measured by reference to a chi-square table or chart (see page 561). As has been indicated in Example 17·5, the statistic chi square (described in Chapter XX) is available as $N\phi^2$. Its degrees of freedom are $(m_c - 1)(m_r - 1)$ when m_c and m_r are the number of columns and rows, respectively. The computed chi square is

---

[1] For a revised form see Yule, *An Introduction to the Theory of Statistics* (1937), pages 68-72.

EXAMPLE 17·5

MEAN-SQUARE CONTINGENCY

Data: Tabulated preliminary (Y) and final (X) estimates of 48 workers according to the categories, good (G), medium (M), and poor (P).

Final Estimate (X)

Preliminary Estimate (Y)		G	M	P	f_r
	G	10	2	1	13
	M	4	17	3	24
	P	2	3	6	11
	f_c	16	22	10	$48 = N$

$f^2/(f_c \times f_r)$, by cells [1]

$$\frac{(10)^2}{(16)(13)} \qquad \frac{2^2}{(22)(13)} \qquad \frac{1^2}{(10)(13)}$$

$$\frac{4^2}{(16)(24)} \qquad \frac{(17)^2}{(22)(24)} \qquad \frac{3^2}{(10)(24)}$$

$$\frac{2^2}{(16)(11)} \qquad \frac{3^2}{(22)(11)} \qquad \frac{6^2}{(10)(11)}$$

Same, reduced to decimal form

0.481	0.014	0.008
0.042	0.547	0.038
0.023	0.037	0.327
0.546	0.598	0.373

$\Sigma = 1.517$

$\phi^2 = 0.517$

$$CC = \sqrt{\frac{\phi^2}{\phi^2 + 1}} = \sqrt{\frac{0.517}{1.517}} = \sqrt{0.3408} = 0.584$$

$$\chi^2 = N\phi^2 = 48 \times 0.517 = 24.816$$

Degrees of freedom $= (m_c - 1)(m_r - 1) = 4$
Tabular 1% level $\chi^2 = 13.3$

[1] The computation may be abbreviated by totaling each row as $\Sigma(f^2/f_c) \div f_r$.

above the 1 per cent chance level, hence may be regarded as highly significant.

Graphic correlation methods.—It has been observed that curvilinear correlation, particularly if it involves multiple relationships, requires extended calculation, for which reason several methods of short-cut graphic approximations have been developed. Besides the reduction of calculation, the methods have the added advantage that they are not restricted to parabolic or other simple mathematical regression curves but may effect more complicated representations of the covariation. This feature may, however, prove to be a disadvantage rather than an advantage, because complicated regressions may be simply adaptations to the peculiarities of the data rather than expressions of any rule of covariation, a consideration that must be taken into account in connection with the correlation ratio as well.

The method of graphic correlation lacks the exact mathematical precision of simpler types of analysis or of that attainable by more complicated approaches to curvilinear relationships, but it has distinct advantages. Frequently, the results closely approximate those attained by more extended mathematical procedures. Usually it is not necessary to proceed further than second approximations, and the process thus represents a means of shortening and simplifying correlation analysis. Most important, the method has the advantage of reflecting, by its emphasis upon the approximate nature of its conclusions, the essential limitation of all correlation analysis, i.e., the fact that all such conclusions should be regarded as approximations. This characteristic is frequently obscured in the involved mathematical manipulations of more complicated procedures.[1]

---

[1] For detailed discussions of the methods of graphic analysis, see L. H. Bean, "Graphic Curvilinear Correlation," *Journal of the American Statistical Association,* 24, December, 1929, pp. 386–398, and "Applications of a Simplified Method of Curvilinear Correlation," United States Department of Agriculture, *Bulletin,* April, 1929. See, also, Mordecai Ezekiel, *Methods of Correlation Analysis,* New York, John Wiley & Sons, Chapters 14 and 15. Also see Appendix, pages 551–553.

READINGS

(See also special and general references, pages 591 and 597.)

BAKER, G. H., "Transformation of Non-Normal Frequency Distributions with Normal Distributions," *Annals of Mathematical Statistics*, V (2), May, 1934, pp. 113–124.

BEAN, LOUIS H., "Application of a Simplified Method of Correlation to Problems in Acreage and Yield Variations," *Journal of the American Statistical Association*, 25 (172), December, 1930, pp. 428–439.

—— " Simplified Method of Graphic Curvilinear Correlation," *Journal of the American Statistical Association*, 24 (168), December, 1929, pp. 386–397.

——"Graphic Curvilinear Correlation," *Journal of the American Statistical Association*, 24, December, 1929, pp. 386–398, and "Applications of a Simplified Method of Curvilinear Correlation," United States Department of Agriculture, *Bulletin*, April, 1929.

BRUCE, DONALD, and REINEKE, L. H., "Correlation Alinement Charts in Forest Research," United States Department of Agriculture, *Technical Bulletin* 210, February, 1931.

DVORAK, AUGUST, "A Simplified Computation of Non-Linear Correlation," *Journal of Educational Research*, 25 (2), February, 1932, pp. 99–104.

EZEKIEL, MORDECAI, "Further Remarks on the Graphic Method of Correlation," *Journal of the American Statistical Association*, 27 (178), June, 1932, pp. 183–185.

——*Methods of Correlation Analysis*, New York, John Wiley & Sons, Chapters 14 and 15.

EZEKIEL, MORDECAI, and INGRAHAM, MARK H., "The Application of the Theory of Error to Multiple and Curvilinear Correlations," *Proceedings of the American Statistical Association*, 24´(165A–Supplement), March, 1929, pp. 99–107.

HARRIS, J. ARTHUR, and CHI TU, "A Second Category of Limitations in the Applicability of the Contingency Coefficient," *Journal of the American Statistical Association*, 24 (168), December, 1929, pp. 367–375.

LI, CHEN-NAN, "Summation Method of Fitting Parabolic Curves and Calculating Linear and Curvilinear Correlation Coefficients on a Scatter Diagram," *Journal of the American Statistical Association*, 29 (188), December, 1934, pp. 405–410.

PEARSON, KARL; HARRIS, J. ARTHUR; TRELOAR, ALLAN E.; and WILDER, MARIAN, "On the Theory of Contingency," *Journal of the American Statistical Association*, 25 (171), September, 1930, pp. 320–327.

WAITE, WARREN C., "Some Characteristics of the Graphic Method of Correlation," *Journal of the American Statistical Association*, 27 (177), March, 1932, pp. 68–70.

WICKSELL, S. D., "Remarks on Regression," *Annals of Mathematical Statistics*, 1 (1) February, 1930, pp. 3–13.

WOO, T. L., "Tables for Ascertaining the Significance or Non-significance of Association Measured by the Correlation Ratio," *Biometrica*, 21 (1–4), December, 1929, pp. 1–66.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Calculate the eta coefficient for the following simplified correlation tables:

(a)

Y \ X	2	4	6	8
12				1
9		2	4	1
6	2	3	1	
3	1	1		

(b)

Y \ X	0	1	2	3	4
4		1	6	1	
3		3	10	5	
2		4	4	3	1
1	1	2		1	4
0	4				

(c)

Y \ X	0	1	2	3	4
4		1	2	3	
3		2	7	3	1
2	3	4	0	1	2
1	2	2	1	1	1
0	3	1			

(d)

Y \ X	0	1	2	3
3		1	1	
2		2	2	1
1	2	2	1	1
0	2			1

(e)

Y \ X	0	1	2	3	4
3			1	1	
2		1	3	2	1
1		0	1	1	1
0	2	1			1

2. Assuming parabolic relationship, compute the regression equation and the correlation index for each pair of correlative series below.

(a)		(b)		(c)	
X	Y	X	Y	X	Y
7	15	5	3	16	12
9	18	11	12	7	10
3	-1	8	16	6	2
15	7	1	1	12	26
6	9	13	6	17	8
		10	16	11	24
				15	16

3. Assuming parabolic relationship, compute the regression equation and the correlation index for each set of data in Exercise 1, above.

4. From the following data, compute both eta and rho (parabolic). Why are they the same?

(a)

Y \ X	0	1	2
6		1	
4		2	2
2	1	1	2
0	1		

(b)

Y \ X	0	1	2
6		2	
4	1	2	
2	2		1
0	1		1

ANSWERS TO EXERCISES

1. (a) Eta = 0.719; (b) eta = 0.785; (c) eta = 0.639; (d) eta = 0.658; (e) eta = 0.706.

2. (a) $T = -17.884 + 7.106X - 0.362X^2$. $\rho^2 = 0.9418$; $\rho = 0.9704$. Least significant index is 0.975.

 (b) $T = -5.06638 + 4.12979X - 0.23715X^2$. $\rho^2 = 0.6390$; $\rho = 0.7993$. Least significant index is 0.930.

 (c) $T = -67.23884 + 15.60621X - 0.66272X^2$. $\rho^2 = 0.9814$; $\rho = 0.9907$. Least highly significant index is 0.949.

3. (a) $T = 3.74999 + 0.54033X + 0.03831X^2$. $\rho = 0.7190$.

   (b) $T = 0.25185 + 2.60926X - 0.59815X^2$. $\rho = 0.7836$.

   (c) $T = 0.98515 + 1.60171X - 0.32541X^2$. $\rho = 0.6063$.

   (d) $T = 0.49461 + 1.89208X - 0.57374X^2$. $\rho = 0.6577$.

   (e) $T = -0.13546 + 1.80513X - 0.37564X^2$. $\rho = 0.6935$.

4. (a) $\eta^2 = \rho^2 = 0.4615$; sq. rt. $= 0.6793$. (b) $\eta^2 = \rho^2 = 0.6667$; sq. rt. $= 0.8165$.

B. PROBLEMS

5. The following data, simplified for purposes of illustration, present results of mental tests (abstract reasoning, X) and efficiency measures (sales, Y). Assuming a parabolic relationship, compute the regression equation and the index of correlation.

Salesman	X	Y
A	2	5
B	3	4
C	3	5
D	9	6
E	3	9
F	6	10
G	4	9
H	7	12
I	5	11
J	8	9

6. From the following assumed data of production and net profit in Factory X, compute the index of correlation, assuming the relationship to be parabolic. State the regression equation.

(a) Production (1000 units)	(a) Profit (100 dollars)	(b) Production (1000 units)	(b) Profit (100 dollars)
1	170	1	106
2	194	2	110
3	196	3	115
4	206	4	116
5	184	5	108

7. Assuming that the following double-frequency table presents comparable records of 50 small independent factories in respect to production (X, units in thousands) and net profits (Y, thousands of dollars), calculate the regression of profits on production, on the hypothesis that the relationship is parabolic. Find the index of correlation.   Is it significant?

Y \ X	3	5	7	9	11
5		1	6	1	
4		3	10	5	
3		4	4	3	1
2	1	2		1	4
1	4				

8. (*a*) Discover the index of correlation for X_1 and X_0 in Exercise 3, Chapter XVI, page 394.   How does this measure compare with the coefficient r?

(*b*) Fit a curvilinear regression representing the ρ just calculated, and compare it with the linear regression.

(*c*) Measure the standard error of the index and compare it with that of the comparable coefficient.

(*d*) Determine the statistical reliability of the index and coefficient by reference to the chart on page 559.

(*e*) Calculate the multiple index $P_{0.123}$, for the data of Exercise 3, Chapter XVI.   How does it compare with the multiple coefficient $R_{0.123}$?   How do they compare as to reliability as measured by the chart?

(*f*) Calculate the multiple curvilinear regression equation for $P_{0.123}$; an estimate X_0 if $X_1 = 15$, $X_2 = 30$, and $X_3 = 20$.

9. (*a*) Fit a curvilinear regression (second degree parabola) to the tabulated data of Problems 10, Chapter XV (page 369).   What are the values of a, b, and c?   Chart the regression with the linear trend already determined.

(*b*) What is the index of correlation?   How does it compare with the coefficient previously obtained?

(*c*) Compare the standard error of this index with that of the comparable coefficient.

(*d*) Compare the reliability of the two measures of covariation by reference to the chart on page 559.

(*e*) Estimate the average sale (Y in dollars) for a store having a capital stock valued at $8,000 ($X$ in thousands) and the limits within which such an average sale would almost certainly fall.

(*f*) Compare the standard errors of estimate of linear and parabolic regressions.

10. To the following data of rainfall and average crop yields fit a parabolic regression and calculate ρ_{yx}. Is the coefficient significant?

County	X Rain per year, inches	Y Yield per acre
A	10	20
B	20	32
C	30	33
D	40	38
E	50	27

11. A parabola fitted to coded data is defined by the equation

$$T = 10 + 2X - X^2$$

The coding is so arranged that the Y taken as origin is $R_y = 4$, and the X taken as origin is $R_x = 15$. The coding is expressed in unit class intervals but the class intervals of the actual data are $i_y = 2$ and $i_x = 5$.

(a) Revise the regression equation to make it applicable to the actual scales.

(b) By means of the regression equation predict values of Y from the following values of X: 5, 15, and 25 ($T = 8, 24, 24$).

(c) Plot the regression equation against the coded X values, $-1, -2, 0, 1, 2$, and check by also calculating these values from the actual data by means of the revised regression equation.

12. The data below, from page 44, are used to represent the records of a group of factory workers in respect to (A) scores in psychological tests designed to measure ability in factory work, and (B) ratings of efficiency as determined in actual work. The two indexes may be plotted as a double-frequency tabulation showing the regression of efficiency on test scores.

(a) From the double-frequency distribution thus tabulated compute the correlation ratio.

(b) Compare the correlation ratio with the corresponding Pearsonian coefficient of correlation (see page 424), and from a comparison of the two determine the linearity of the distribution.

(c) Compute the index of correlation, and explain its relation to r for the following data:

	Index A	Index B		Index A	Index B
A	13	19	I	22	24
B	17	16	J	14	16
C	13	15	K	9	14
D	19	25	L	9	11
E	14	8	M	14	14
F	10	6	N	6	4
G	· 11	12	O	19	21
H	19	20	P	15	15.

13. Measure the data of Problem 12, Chapter XV, by calculating the coefficient of mean square contingency and appraising its significance.

14. Raw materials for a certain manufacturing concern are available in three grades, which are designated a, b, and c, of which c is lowest. Question is raised as to whether there is a significant difference in the proportions of finished products derived from each grade that pass inspection. The results of a recent run may be summarized as follows:

Of 510 units from grade a material, 371 passed, 139 were rejected,
Of 580 units from grade b material, 243 passed, 337 were rejected,
Of 540 units from grade c material, 127 passed, 413 were rejected.

Appraise this comparison by calculating the coefficient of mean square contingency, and chi square as $N\phi^2$.

CHAPTER XVIII

THE CORRELATION OF TIME SERIES

Many important applications of the principles of correlation appear in the analysis of time series, particularly in studies pertinent to the business cycle. While the general principles of correlation applied in such cases are the same as those explained in earlier chapters, certain problems are of particular significance in connection with time series. One such problem is the distinguishing between seasonal, cyclical, and trend influences. Another is concerned with the tendency of some series to *lead* or precede, whereas others *lag*. In other words, allowance must be made for the fact that cyclic changes are not necessarily concurrent, that the change in some series precedes that in others. These problems will be given special consideration in this chapter.

Trend and cycle.—When time series are to be correlated, it is necessary first to define clearly the objective of the inquiry. It may, for example, be of interest to know how one set of data, such as a monthly index of manufacturing production, is related to a corresponding index of mineral production. In such a case, as a preliminary expedient, the two series might be set up as X and Y, respectively, and the correlation—but not the significance—measured in the manner described in Chapters XIV and XV. Such a correlation would obviously measure the tendency of changes in one series to conform to changes in the other series. But unless the seasonal factor, for example, had been removed, it might obscure a considerable measure of correlation in the two series.

Further, it might be possible that two series would show some correlation with respect to the cycle and not to the trend. For example, over a comparatively normal period of several decades the trend of industrial production may have been

436

upward while that of the interest rate was falling. Yet both series might be subject to similar cyclic fluctuation. If an ordinary correlation were made between two such series, the trend effect would probably predominate, but would to some extent be offset by the cyclic factor. Under these conditions it would be desirable to eliminate the trend, the correlation of which would be fairly obvious in any case, and then to correlate the deviations representing the respective cyclical fluctuations. The result obviously would not show the relation between production and interest rates as such but rather the relation of the two series with respect to their tendency to deviate from their trend. Most correlations of time series involve this problem of separating various factors or influences so that their specific covariation may be measured.

The correlation of cycles.—In Example 18·1 a very abbreviated illustration is presented of the correlation of cyclic change in industrial production with that in wholesale prices for the years 1923–1929, inclusive. An inspection of the problem will indicate that the first step is the calculation of the trend. A least-squares straight-line trend has been chosen as appropriate, but in problems involving longer series of data a parabolic trend, a modified geometric, or possibly a moving average might be more appropriate. In other words, the first problem encountered is discovering and calculating the type of trend appropriate to the given series. If the data are quarterly or monthly, the trend, as explained in Chapter XIII, would first be fitted to the annual averages, and then broken down to quarterly or monthly items, corresponding with the data.

After an appropriate trend has been calculated, it may be removed from the data, preferably by division, but often more conveniently by subtraction. In the problem at hand, both methods ($100 Y/T - 100$ and $Y - T$) have been employed. However, in a long time series, where the trend changes markedly from low values to high values, the percentage measure should generally be used, since it tends to equalize the average extent of the deviations on the different trend levels.

After the deviations from normal (trend) have thus been discovered, the measurement of correlation takes the same form

EXAMPLE 18·1

CORRELATION OF TIME SERIES

Data: Indexes of industrial production (Q) and wholesale prices (P) in the United States, 1923–1929.

Indexes expressed as percentage deviations (d per cent), from a straight-line trend, T, and in difference deviations (d).

	Year	Index Y	Trend T	Cyclical percentage, C per cent $=Y \div T$	Deviation percentage, d per cent $=C$ per cent-100	Difference deviations $d=Y-T$
	1923	101	96.86	104.27	4.27	4.14
	1924	95	100.00	95.00	−5.00	−5.00
	1925	104	103.14	100.83	0.83	0.86
	1926	108	106.29	101.61	1.61	1.71
Industrial	1927	106	109.43	96.87	−3.13	−3.43
production	1928	111	112.57	98.61	−1.39	−1.57
	1929	119	115.71	102.84	2.84	3.29
		744	744.00	700.03	0.03	0.00
					$\sigma=3.0778$	$\sigma=3.1816$
	1923	100.6	101.39	99.22	−0.78	−0.79
	1924	98.1	100.43	97.68	−2.32	−2.33
	1925	103.5	99.47	104.05	4.05	4.03
Wholesale	1926	100.0	98.51	101.51	1.51	1.49
prices	1927	95.4	97.56	97.79	−2.21	−2.16
	1928	96.7	96.60	100.10	0.10	0.10
	1929	95.3	95.64	99.64	−0.36	−0.34
		689.6	689.60	699.99	−0.01	0.00
					$\sigma=2.0597$	$\sigma=2.0461$

Correlation:

In percentage deviations:

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{19.8179}{7 \times 3.0778 \times 2.0597} = 0.447$$

In difference deviations:

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{20.5263}{7 \times 3.1816 \times 2.0461} = 0.450$$

as in earlier examples. As a rule, such rectilinear correlation is appropriate. It will usually involve the calculation of the simple r only, inasmuch as any curvilinear effect that might be present will probably have been eliminated in the trend. Hence all that is required is the calculation of either the sums of the squares or the standard deviation of each series, together with the cross products. The values thus obtained may be substituted in an equation for r and the coefficient thus determined. In the illustrative problem, the coefficient is $r = 0.447$, as based upon the percentage deviations. This measure varies only slightly when it is based upon the difference deviations.[1]

Reliability in time series.—The problem of determining significance in time-series correlations is somewhat more complicated than for other data. One difficulty arises from the fact that successive items in a time series are themselves more or less linked or correlated, whereas the theory on which the usual measurement of reliability is based assumes discrete measurements, such as are obtained in laboratory experiments by successive throws of coins or dice, or chance drawings from an urn. It is obvious that industrial production of one year is not something independent of that of the previous year, but is, rather, the level established in the preceding year modified by certain new or changing factors. Hence, most authorities agree that no attempt should be made to measure reliability by F, or otherwise, in a direct correlation of two time series. Some argue, however, that the reliability of correlation of two cycles may be so measured whenever deviations from trend approximate normal distributions. Then the cycles are considered direct measures of certain variable forces, and the correlation

[1] If cycles have been expressed in average deviations (d/AD), a so-called coefficient of similarity (Sm) may readily be obtained. From each d/AD pair, the numerically smaller is selected and written with the correlation sign (like signs give plus; unlike, minus). The algebraic average of these items is the coefficient Sm. Normally, its relation to r is $r^2 = 2Sm^2 - Sm^4$. An illustration appears in the Appendix, pages 553–554 (see "First Moment Correlation," by G. R. Davies, *Journal of the American Statistical Association*, December, 1930, and *Methods of Statistical Analysis*, by Davies and Crowder, John Wiley & Sons, 1933). This method is a convenient measure of the relationship between two cycles and other correlative series, and is free from the assumption of normal distributions implied in a correlation surface. Of course, r may be utilized as purely a descriptive measure aside from these assumptions.

indicates the degree to which one series covaries with another, above what might be expected by chance.

There are, of course, time series in which the fact of linkage does not seriously interfere with the measure of significance. For example, if a workman is tested with respect to his efficiency in operating a machine, the interval between the tests being comparatively long, results are likely to be obtained which represent fairly well the normal variability of the individual under diverse conditions of biological and environmental factors. If, at the same time, the workman is improving through practice, this improvement will show itself as a rising trend on which the chance fluctuations in efficiency are superimposed. In such a case, the significance of the trend, that is, the correlation of the data with time, may be of importance in evaluating proof of increasing efficiency.

Prediction in time series.—Cases may arise where it seems desirable to predict the value of one time series from that of another series. This is most likely to occur when one series is reported more promptly than the other, and an estimate of the second is required. Assuming that the series in question exhibit fairly constant trends, it would be permissible to extrapolate both trends a single unit to the current date. On the basis of the deviation of the reported series (x), the deviation (y) from the other trend could then be estimated from the usual regression equation derived from the correlation of the cycles, although in this case, since each set of deviations is practically centered, the regression equation will take the form

$$t \text{ or } y' = bx$$

where x is the current deviation from trend of the series reported, y is the estimated deviation, and b, as usual, is $\Sigma xy/\Sigma x^2$, calculated from deviations from the respective trends. The y' thus estimated may be added algebraically to the extrapolated trend item to obtain the required estimate, or $1 + y'$ may be multiplied by T if the percentage form has been employed. The reliability of such an estimate, of course, is difficult to determine, since it depends not only on the degree of correlation and the probability of the correlation's holding constant over a

period of time, but also upon the probability that the trend will remain consistent. Whether such an estimated figure is to be used or not is largely a matter of judgment and experience rather than of a calculated measure of reliability.

The use of partial correlation.—In measuring the degree of cyclic correlation, it is sometimes convenient to make use of partial correlation as a means of eliminating the trends. This procedure is illustrated in Example 18·2, where assumed index numbers of production and prices are compared. The production trend is distinctly upward, while the price trend is downward. The objective is to measure the correlation of cyclic change in the two series.

The problem might, of course, be approached by the method presented in Example 18·1. That is, trends might be fitted to each series, and the x_0 deviations $(X_0 - T_0)$ correlated with the x_1 deviations $(X_1 - T_1)$. The same measure of correlation, however, may be obtained more directly by means of a partial correlation in which X_1 is the index of production, X_2 is a straight-line trend written in any convenient form as 1, 2, 3, ... N, and X_0 is the price series. The calculation of the centered squares and cross products (Np) is shown in the example, together with the computation of the multiple regression equation and the measures of multiple and partial correlation. As a check on the work, predictions of each X are obtained on the basis of the regression equation. Data and trends are illustrated in Fig. 18·1.

As a further check, the coefficient $r_{01.2}$ may be obtained by the difference method of Example 18·1. It will be found that the required trends are

$$T_1: \quad 15, \ 19, \ 23, \ 27, \ 31, \ 35, \ 39, \ 43, \ 47, \ 51$$
$$T_0: \quad 40, \ 38, \ 36, \ 34, \ 32, \ 30, \ 28, \ 26, \ 24, \ 22$$

and the deviations from trend are

$$d_1 \text{ or } x: \quad 2, \ -4, \ -1, \ 0, \ 2, \ 3, \ 1, \ -1, \ -2, \ 0$$
$$d_0 \text{ or } y: \quad 2, \ -4, \ -5, \ 4, \ 3, \ 3, \ 1, \ -1, \ -3, \ 0$$

The squares and cross products are

$$\Sigma x^2 = 40; \quad \Sigma y^2 = 90; \quad \Sigma xy = 48$$

Example 18·2

CORRELATION OF CYCLES BY PARTIAL CORRELATION

Data: Assumed figures of production (X_1, thousands of tons) and price (X_0, dollars per ton) for a certain industry, 1931–1940.

Year	X_1	X_2	X_0	Z	X_0'	d
1931	17	1	42	60	42.4	−0.4
1932	15	2	34	51	33.2	0.8
1933	22	3	31	56	34.8	−3.8
1934	27	4	38	69	34.0	4.0
1935	33	5	35	73	34.4	0.6
1936	38	6	33	77	33.6	−0.6
1937	40	7	29	76	29.2	−0.2
1938	42	8	25	75	24.8	0.2
1939	45	9	21	75	21.6	−0.6
1940	51	10	22	83	22.0	0.0
S	330	55	310	695	310.0	0.0
						$\Sigma d^2 = 32.4$

P	1	12,250	2,145	9,618			
	2		385	1,540			
	0			10,030	49,271	check	

Np	1	13,600	3,300	−6,120			
	2		825	−1,650			
	0			4,200	9,685	check	

$$T = 28.8 + 1.2X_1 - 6.8X_2$$
$$R_{0.12} = 0.9607$$
$$r_{01} = 0.8098$$
$$r_{01.2}^2 = \frac{(\Sigma x_2^2 \Sigma x_1 x_0 - \Sigma x_2 x_0 \Sigma x_1 x_2)^2}{\left(\Sigma x_1^2 \Sigma x_2^2 - \overline{\Sigma x_1 x_2}^2\right)\left(\Sigma x_2^2 \Sigma x_0^2 - \overline{\Sigma x_2 x_0}^2\right)}$$
$$= \frac{[(-825 \times 6,120) + (1,650 \times 3,300)]^2}{[(13,600 \times 825) - 3,300^2)][(825 \times 4,200) - 1,650^2)]}$$
$$= \frac{396,000^2}{330,000 \times 742,500} = 0.64$$
$$r_{01.2} = 0.80$$

Hence the coefficient measuring the correlation of the cycles is

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{48}{\sqrt{40 \times 90}} = 0.80$$

which is necessarily the same as that obtained by the partial correlation just computed.
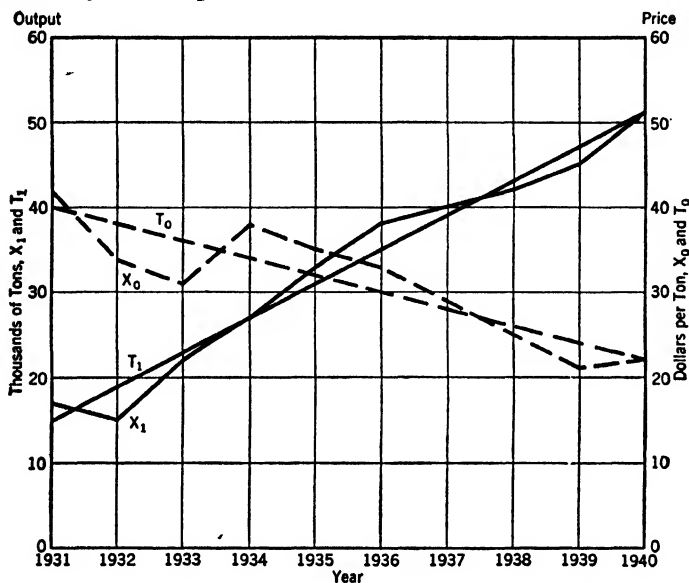
FIG. 18·1.—Trends of Production and Prices. Data: Assumed production and price of output of a certain industry, 1931–1940 (see Example 18·2).

Prediction by multiple correlation.—The method of predicting a new X_0 item from an X_1 reported item has already been described. However, where difference deviations ($d = Y - T$) are satisfactory, prediction may more conveniently be made by the use of a multiple regression equation made up from the same centered squares and cross products (Np) that have been calculated in the measurement of the partial correlation coefficient. When this equation has been obtained, the values of X_1 and X_2 may be substituted to obtain an estimate of X_0. To illustrate, in Example 18·2 let us assume that in 1941 a figure of 57 is reported for X_1, and an estimate of X_0 is required. If it may be assumed that the established trends are still in

force, the required estimate may readily be obtained by substituting $X_1 = 57$ and $X_2 = 11$ in the regression equation thus:

$$T = 28.8 + (1.2 \times 57) - (6.8 \times 11) = 22.4$$

While such predictions obviously could not be relied upon except where the extrapolation is small and the situation is undisturbed, nevertheless the method has its uses. As a rule, however, it would be desirable to recompute the regression equation for each prediction so as to minimize the degree of extrapolation.

Allowance for lag.—When two time series in either their original form or expressed as cycles are to be correlated, a comparison will often show that the cycle of one series leads while the other lags. For example, the cycle of stock prices on the New York Stock Exchange formerly preceded the cycle of industrial production by about 6 months. In such a case, the correlation may be worked out in the usual manner, except that the X and Y series should be so placed that the cycles rather than the months concur. To illustrate, the stock-market figure for December, 1907, taken as X might be paired with the industrial production figure for June, 1908, taken as Y, and so on for succeeding items. The degree of lag may be estimated by plotting the cycles to the same scale on transparent paper, superposing one on the other, and noting the amount of lag when the cycles appear most closely to coincide. A more exact procedure requires calculating the correlation at the estimated lag and also at several points chosen with a shorter and longer lag. The point at which the correlation is highest will determine the lag to the nearest time unit employed.

A distributed lag.—Professor Irving Fisher has shown that the effect of a lag may not be carried forward with any exact time interval, but may tend to be distributed. For example, suppose it to be assumed that business activity as reflected in bank debits stimulates consumption 6 months later. Under these conditions it would not be expected that the full influence would be felt exactly 6 months later, but rather that it would spread like a wave over several months. That is, on the average its influence might be felt at the stated interval, but it

would also affect earlier and later months to some extent. In other words, the lag effect, or perhaps merely the lag relation, would be distributed.

At first Professor Fisher assumed that the wavelike distributed lag would take the form of a positively skewed distribution with the "tail" of the distribution extending with diminishing force over many succeeding months. But such a lag is difficult to compute. Also, in practice it was found that often, particularly where the lag was short, a wave with its mode in the next month, and its influence waning in succeeding months, represented a satisfactory approximation. That is, the total influence of the *leading* (i.e., preceding) series, if distributed over the succeeding 3 months in the lagging series, is assumed to affect these months in the ratio 3, 2, and 1. Or a 6-month distributed lag would be assumed to distribute its influence in the ratio 6, 5, 4, 3, 2, and 1 over the succeeding 6 months.

The computation of such a lag—or rather the projecting forward of the leading series to neutralize the lag—is relatively easy. The procedure is based on the principle that in any month the influence of the leading series will be exerted in terms of a distribution like that described, but in reverse order. For example, in the month of July the influence of a 6-month distributed lag would be received in the ratio:

6 parts from June
5 parts from May
4 parts from April
3 parts from March
2 parts from February
1 part from January

In accordance with the foregoing principles the relative force of the lag just described, as registered in July, could be calculated as an average of the January–June items weighted successively by 1, 2, 3, 4, 5, and 6. Thus, the weighted total is (preceding items $= X_p$; weights $= w$):

MONTH:	J	F	M	A	M	J	Σ
X_p:	2	1	3	2	4	6	18
w:	1	2	3	4	5	6	21
wX_p:	2	2	9	8	20	36	77

The same computation may then be made for August, thus:

MONTH:	F	M	A	M	J	J	Σ
X_p:	1	3	2	4	6	4	20
w:	1	2	3	4	5	6	21
wX_p:	1	6	6	16	30	24	83

The weighted averages may be found as $\Sigma wX_p/\Sigma w$, but, as the lagged influence is measured only relatively and not absolutely, the weighted sums will do just as well. Hence the procedure just indicated, in which projected effects are measured as 77 for July and 83 for August, could be continued throughout the series, and the results could be correlated with the lagging series (Y), assuming that the average amount of lag is correct. It should be noted that the *average* lag is not 6 months, but only 2.67 months.

Short-cut distributed lag.—Like the moving average, the distributed lag may be computed by a short-cut method, as indicated in Example 18·3. In that example the leading (i.e., preceding) series is designated X_p, and the lagging series, Y. The allowance for an average 2.67-month distributed lag, as described above, is made by first computing a moving sum ($M\Sigma$) of 6, or, in general, n items, the sum being entered opposite the seventh item. This moving sum is the same as that computed for an n-term moving average, except that it is entered in the row *following* the items added. The next item is the sum of the second to the seventh items, inclusive, and it is entered in the eighth row. It may obviously be found as $18 - 2 + 4 = 20$. Similarly, the next is $20 - 1 + 1 = 20$, etc.

The first X (i.e., 77) is computed as a weighted average of the first 6 items (weights 1, 2, 3, 4, 5, and 6) as previously explained, entered in the seventh row. The computation of the next X (i.e., X_{+1}) may be short-cut, as

$$X_{+1} = X - M\Sigma + nX_p$$

The items required for this calculation are contained in the seventh row, and the result, $X = 83$, is entered in the eighth row. In the same manner succeeding X's may be found. The

correlation between X and Y ($N = 18$), July, 1939, to December, 1940, may then be calculated by one of the methods described in earlier chapters. The series are, of course, too short for satisfactory correlation of cycles. As has previously been explained, reliability must be judged by experience with the data correlated, rather than by probability methods.

EXAMPLE 18·3

COMPUTATION OF A DISTRIBUTED LAG

Data: Simplified monthly cycle series X and Y, the latter lagging about 3 months (distributed lag over $n = 6$ months, or average lag of 2.67 months).

		X_p	$M\Sigma$	X	$M\Sigma +$	$(n \times X_p)$	$= X_{+1}$	Y
1939	J	2						3
	F	1						5
	M	3						7
	A	2						6
	M	4						5
	J	6						3
	J	4	18	77	$- 18 +$	6×4	$= 83$	7
	A	1	20	83	$- 20 +$	6×1	$= 69$	9
	S	1	20	69	$- 20 +$	6×1	$= 55$	6
	O	2	ʹ18	55	$- 18 +$	6×2	$= 49$	3
	N	3	18	49	$- 18 +$	6×3	$= 49$	2
	D	5	17	49	$- 17 +$	6×5	$= 62$	3
1940	J	4	16	62		etc.		4
	F	3	16	70				5
	M	2	18	72				7
	A	3	19	66				5
	M	2	20	65				5
	J	5	19	57				3
	J	3	19	68				6
	A	4	18	67				5
	S	3	19	73				7
	O	2	20	72				6
	N	1	19	64				4
	D	2	18	51				3

Modified form of distributed lag.—The calculation of a distributed lag, just described, may be readily adapted to other forms of distribution. Suppose, for example, that the forward

distribution of each item appeared to be more plausibly represented, during succeeding months, by the ratio [1]

Month: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
Weight: 2, 4, 6, 8, 7, 6, 5, 4, 3, 2, 1 $\Sigma w = 48$

thus setting the mode in the fourth month and extending the range over 11 months. As before, this set of ratios may be applied in reverse order to preceding months in order to determine the accumulated influence exerted in any given month. That is, the total influence projected forward (X) to December from prior months of the same year, assuming the data of Example 18·3, would be:

Mo.:	J	F	M	A	M	J	J	A	S	O	N	Σ
X_p:	2,	1,	3,	2,	4,	6,	4,	1,	1,	2,	3	
w:	1,	2,	3,	4,	5,	6,	7,	8,	6,	4,	2	48
wX_p:	2,	2,	9,	8,	20,	36,	28,	8,	6,	8,	6	133

The total (X) for the next month, January, could be obtained directly by moving the weights forward a month, or by entering for December an 8-term moving total of the January–August items $(M\Sigma_1)$, and a 4-term moving total of the September–December items $(M\Sigma_2)$ and taking the January total $(X_{\text{Jan.}})$ as

$$X_{\text{Jan.}} = X_{\text{Dec.}} - M\Sigma_1 + 2M\Sigma_2$$

or, in general,

$$X_1 = X - M\Sigma_1 + 2M\Sigma_2$$

$$= 133 - 23 + 2 \times 11 = 132$$

The form may be set up paralleling Example 13·6, and the required items yielding the January total will be in the Decem-

---

[1] Many other combinations of weights could be made, approximating a logarithmic normal, for example:

at $X_{\text{Dec.}}$ w: =	1,	1,	1,	2,	2,	2,	3,	3,	4,	2,	1,	0
at $X_{\text{Jan.}}$ w:	0,	1,	1,	1,	2,	2,	2,	3,	3,	4,	2,	1
Change:	−1,	0,	0,	−1,	0,	0,	−1,	0,	−1,	+2,	+1,	+1

In such a case the transition from the December total $(X_{\text{Dec.}})$ to the next month's total $(X_{\text{Jan.}})$ could be made most conveniently by noting the change in the weight. These changes may be indicated on the edge of a card, which is placed opposite the requisite X_p items, and the calculation carried on a machine.

ber row. Obviously, the $M\Sigma$'s and the calculations may be continued to the end of the series. The reliability of correlations thus calculated cannot, of course, be determined by probability methods but must be judged by experience with the situations in which they occur.

READINGS

(Also see special and general references, pages 591 and 597.)

BLACKETT, O. W., and WILSON, W. P., "A Method of Isolating Sinusoidal Components in Economic Time Series," *Michigan Business Studies*, University of Michigan, 8 (4), 1938, pp. 1–58.

DAVIES, GEORGE R., "First Moment Correlation," *Journal of the American Statistical Association*, 25 (172), December, 1930, pp. 413–427.

HAFSTAD, L. R., "On the Bartels Technique for Time Series Analysis," *Journal of the American Statistical Association*, 35 (210), June, 1940, pp. 347–361.

KUZNETS, SIMON, "On the Analysis of Time Series," *Journal of the American Statistical Association*, 23 (164), December, 1928, pp. 398–410.

—— "Random Events and Cyclical Oscillations," *Journal of the American Statistical Association*, 24 (167), September, 1929, pp. 258–275.

—— "Seasonal Pattern and Seasonal Amplitude; Measurement of Their Short-time Variations," *Journal of the American Statistical Association*, 27 (177), March, 1932, pp. 9–20.

MACAULAY, FREDERICK, R., "The Smoothing of Time Series," National Bureau of Economic Research, *Monograph* 19, February, 1931.

RASOR, EUGENE, "The Use of the Time Factor as a Variable in Correlation Computations," Ohio State Bureau of Business Research Publication, June, 1930, pp. 25–29.

ROOS, CHARLES F., "Annual Survey of Statistical Techniques: The Correlation and Analysis of Time Series," *Econometrica*, 4 (4), October, 1936, pp. 368–381.

TINBERGEN, J., *A Method and Its Application to Investment Activity* and *Business Cycles in the United States of America, 1919–1932*, Vols. I and II, League of Nations, Geneva, 1939.

TINTNER, GERHARD, "The Analysis of Economic Time Series," *Journal of the American Statistical Association*, 35 (209), March, 1940, pp. 93–100.

—— *The Variate Difference Method*, Bloomington, Indiana, Principia Press, 1940, 167 pp.

YULE, G. U., "On the Time Correlation Problem, with Especial Reference to the Variate Difference Correlation Method," *Journal of the Royal Statistical Society*, July, 1921, pp. 497–537.

—— "Why Do We Sometimes Get Nonsense Correlations Between Time Series?" *Journal of the Royal Statistical Society*, Vol. 89, 1926.

EXERCISES AND PROBLEMS

A. EXERCISES

1. The following paired sets of data (X_1 and X_0) parallel approximately the assumed production and price data of Example 18·2, where X_2 is a slope. By the methods there indicated, calculate $T_{0.12}$, $R_{0.12}$, and $r_{01.2}$, and check the last item by correlating $X_0 - T_0$ and $X_1 - T_1$.

(a)		(b)		(c)		(d)		(e)	
X_1	X_0	X_1	X_0	X_1	X_0	X_1	X_0	X_1	X_0
24	16	22	50	42	16	15	41	17	42
20	14	22	43	34	14	17	36	15	34
25	17	25	41	35	17	22	36	18	35
28	30	29	39	34	30	28	36	31	34
32	33	33	37	34	33	34	36	34	34
35	37	34	33	33	37	37	34	38	33
35	39	34	30	29	39	39	33	40	29
35	41	35	17	25	41	42	22	42	25
36	43	34	14	22	43	43	21	44	22
40	50	42	16	22	50	53	25	51	22

2. The following series X_p are assumed to lead another (lagging) series, not given, with which they are to be correlated. Assuming weights of 1, 2, 3—the weighted total projected to the succeeding year—find the required totals, by use of a projected 3-term moving total; e.g., in (a), $24 + 20 + 25 = 69$ is written under $M\Sigma$ in the fourth year, etc., and $24 + (2 \times 20) + (3 \times 25) = 139$ is X in that year. Then the next X is $139 - 69 + (3 \times 28) = 154$, etc.

(a)	(b)	(c)
X_p	X_p	X_p
24	22	42
20	22	34
25	25	35
28	29	34
32	33	34
35	34	33
35	34	29
35	35	25
36	34	22
40	42	22

1.

	$T_{0.12}$	$R_{0.12}$	$r_{01.2}$
(a)	$-14.00 + 1.20X_1 + 1.60X_2$	0.9884	0.80
(b)	$30.00 + 1.20X_1 - 6.40X_2$	0.9884	0.80
(c)	$-40.40 + 1.20X_1 + 6.40X_2$	0.9884	0.80
(d)	$29.80 + 1.20X_1 - 6.80X_2$	0.9607	0.80
(e)	$36.13 + 0.53X_1 - 4.13X_2$	0.9803	0.80

2.

(a)		(b)		(c)	
$M\Sigma_{(3)}$	X	$M\Sigma_{(3)}$	X	$M\Sigma_{(3)}$	X
69	139	69	141	111	215
73	154	76	159	103	206
85	177	87	182	103	205
95	197	96	197	101	201
102	207	101	203	96	187
105	210	103	207	87	166
106	213	103	206	76	145

B. PROBLEMS

3. Assuming that the following data represent production (X_1), a slope (X_2), and prices (X_0), calculate the relationship between the cycles of production and prices $(r_{01.2})$ and a regression equation $(T_{0.12})$. On the assumption that the trends of (X_1) and (X_0) still obtain, predict X_0 for the year 1926. Check the partial correlation by eliminating straight-line trends from X_1 and X_0 $(X - T)$ and correlating the difference cycles.

YEAR	X_1	X_2	X_0
1919	14	1	25
1920	16	2	19
1921	18	3	19
1922	28	4	23
1923	31	5	20
1924	31	6	18
1925	37	7	16

4. Calculate $r_{01.2}$ from the following data allowing for a lag of about 3 years in the cycle of X_0. Check the projected totals of X_p (parentheses under X_1, 1932–1938), using weights 1, 1, 2, 2, 3, 1 (1926–1931) to obtain the 1932 total of 267, and adding items indicated by weights $-1, 0, -1, 0, -1, 2, 1$ (1926-1932) to obtain 311. Further totals are similarly obtained by additions as indicated by the latter weights moved forward. Find $r_{01.2}$ and the multiple regression equation.

Year	X_p	X_1	X_2	X_0	Z
1926	17			44	
1927	15			45	
1928	22			43	
1929	27			42	
1930	33			34	
1931	38			31	
1932	40	(267)	1	38	306
1933	42	(311)	2	35	348
1934	45	(353)	3	33	389
1935	51	(387)	4	29	420
1936	57	(421)	5	25	451
1937	60	(462)	6	21	489
1938	61	(505)	7	22	534
Total		(2,706)	28	203	2,937

5. (*a*) The two independent series, X_1 and X_2, below, have least-squares linear trends ($b_1 = 1$; $b_2 = 2$), while the dependent series, X_0, has none. Calculate $R^2_{0.12}$.

(*b*) Subtract out the trends from X_1 and X_2, obtaining x_1 and x_2, and recompute the correlation.

(*c*) To the data of (*a*) add X_3, the slope 1, 2, 3, . . . 7, and find $R^2_{0.123}$. Explain why this agrees with the R^2 obtained in (*b*).

(*d*) Compute regression equations for correlations (*b*) and (*c*), and predict regression items (T or X_0) in each case.   Why do they agree?

Year	X_1	X_2	(X_3)	X_0
1934	3	1	(1)	2
1935	8	6	(2)	7
1936	3	4	(3)	3
1937	9	11	(4)	6
1938	6	7	(5)	1
1939	10	13	(6)	5
1940	10	14	(7)	4

6. From the *Survey of Current Business* obtain data of bank debits and department-store sales in the United States for the years 1923–1929, inclusive. Assuming that the cycle of bank debits leads that of department-store sales by approximately 6 months, allow for a distributed lag of the former, employing the following weights, applied to January, 1923, and successive months:

Weights = 1, 1, 2, 2, 3, 3, 4, 4, 5, 4, 3, 2, 1 (Jan., '23 . . . Jan., '24)

The weighted total is entered as of February, 1924. The March total is obtained by adding to the above total the items indicated by the following weights:

$$\text{Weights} = -1, 0, -1, 0, -1, 0, -1, 0, -1 + 1 + 1 + 1 + 1 + 1$$
$$(\text{Jan., '23} \ldots \text{Feb., '24, etc.})$$

Successive totals may be obtained by repeated use of the latter set of weights moved forward one month.

Correlate the weighted totals (X_1) thus projected forward, with the adjusted indexes for department-store sales (X_0) entering

$$X_2 = 1, 2, 3, \ldots N$$

as a second independent series with department-store sales as dependent. Find and interpret the coefficient of partial correlation, $r_{01.2}$.

CHAPTER XIX

THE ANALYSIS OF VARIANCE

In preceding chapters rather extended attention has been given to the measurement of covariation and to methods of appraising the reliability of various measures of correlation. In the present chapter attention turns to what may be described as an extension of the correlation technique. That extension involves a group of statistical procedures whose purpose is generally described as the analysis of variance. These procedures have come into accepted usage only recently and in a limited range of applications, particularly in agricultural statistics, but they have a broad potential usefulness.

The term *analysis of variance* implies a study of problems of variability, particularly as variability is represented by the squared standard deviation, and, in its simplest form, includes such problems as the significance of the difference of means for both uncorrelated and correlated data. An introduction to the subject, therefore, may very well begin with this particular topic.

Difference between two means.—Questions regarding the significance of the difference of means arise typically when comparisons are made between groups, as in the case of average performance for two comparable groups of machine operators, salesmen, or other employees, or by the same group under different conditions. Suppose, for example, that the management in a large firm sought to compare the relative performance of comparable workers in two different factories. Chance variability would be likely to cause some difference between average performance in one shop and that in the second plant. Question would arise, however, whether the actual difference between the two was greater than would be likely to occur by chance, in other words, whether the two levels of performance were so

distinctly different as to preclude the likelihood that their divergence was only accidental. If the difference could not be explained by chance, measures might properly be taken to improve the level of performance in the lagging plant. Questions of this sort are most appropriately answered by the procedures of analysis of variance.

Similarity to correlation analysis.—A preliminary approach to this question may be made through the familiar procedure of correlation, for, as has been indicated, correlation is one form of variance analysis. The use of correlation technique for this purpose may be illustrated by reference to the highly simplified data of Table 19·1. According to the data there presented, the

TABLE 19·1

PERFORMANCE RECORDS OF WORKERS IN TWO PLANTS

Data: Assumed for illustrative purposes.

Plant I		Plant II	
Employee	Performance in units Y_1	Employee	Performance in units Y_2
A	4	F	9
B	2	G	5
C	5	H	8
D	8	I	12
E	6	J	14
—		K ·	12
			—
Total	25		60
M	5		10

workers in group 1 have records (Y_1) averaging 5, and those in group 2 (Y_2) average 10. The question for consideration is whether the difference between these means is to be interpreted as only chance variability or as representative of significantly different levels of performance.

The question may be answered by describing the series of records in the first plant as each having an X value of 0, while those in the second plant each have an X value of 1, thus:

Employee:	A	B	C	D	E	F	G	H	I	J	K
Performance (Y):	4,	2,	5,	8,	6,	9,	5,	8,	12,	14,	12
Plant (X):	0,	0.	0.	0.	0,	1,	1,	1,	1,	1,	1

The X and Y series may then be correlated.   Figure 19·1 shows this arrangement and the regression line, which passes through the mean of each group.   The two X values may, of course, be any two selected figures, hence the use of 0 and 1 for convenience
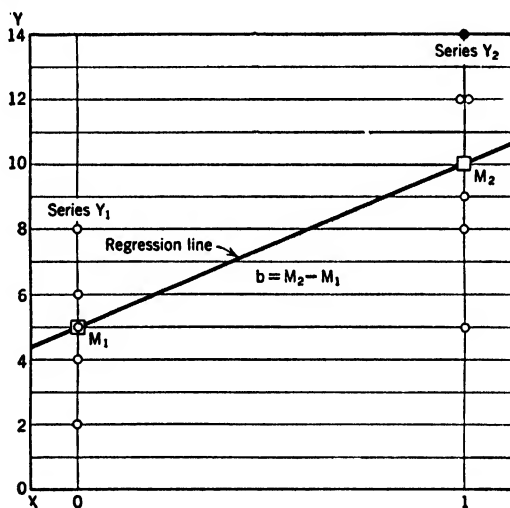
Fig. 19·1.— Significance of the Difference between Two Means Evaluated by Correlation.   Data from Table 19·1.

in calculation.   On the basis of this X scale, it is readily calculated that

$$ r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{13.6364}{\sqrt{2.7273 \times 142.1818}} = 0.6925 $$

Reference to charts A·3 and A·4, pages 559 and 560, indicates that this coefficient is significant, though not highly so (when

$N = 11$, the 5 per cent level is 0.60 and the 1 per cent level is 0.74).

The interpretation is obvious, as a simple comparison will show. It will be recalled that a significant correlation of $r = 0.683$ was found in the covariation of efficiency, as measured by sales, with scores made in a psychological test (cf. Example 14·1, page 327). This indicated that the salesmen who had made high scores proved to be distinctly more efficient than those who had made low scores. If that comparison had been based on psychological tests evaluated merely as high $(X = 1)$ or low $(X = 0)$—thus classifying the salesmen into two contrasting groups—the case at hand would have been a close parallel. The interpretation clearly is that the workmen in plant II are distinctly better than those in plant I, after suitable allowance for performance variabilities is made.

In the analysis of variance, however, significance is generally measured by F, which in this case (linear regression) may be computed as

$$F = \frac{r^2}{1 - r^2} (N - 2) = \frac{0.4796}{0.5204} (11 - 2) = 8.29$$

F measures distinctive dissimilarity, and by reference to the table of F, page 586 (col. 1, row 9), 5 per cent and 1 per cent levels of significance may be found. In this case, the least significant (5 per cent) F value is 5.12, and the least highly significant (1 per cent) F value is 10.56. The value of F found in this example, therefore, is significant but not highly significant. Detailed explanation of the procedure to be followed in using the table of F will be given later in this chapter.

Though this correlation process is somewhat cumbersome and is not easily adaptable to more complex problems, it serves the purpose of presenting the elementary principle of analysis of variance, which involves a comparison of the variability between groups (regression) and within groups (alienation). The process may be simplified algebraically to the following

formula,[1] where the combined sums of squared centered deviations (Σy_1^2 and Σy_2^2) is the correlation term Σd^2:

$$F = (M_2 - M_1)^2 \div \left[\frac{\Sigma y_1^2 + \Sigma y_2^2}{N - 2} \cdot \frac{N}{N_1 N_2} \right]$$

$$= (M_2 - M_1)^2 \div \frac{N \Sigma d^2}{N_1 N_2 (N - 2)}$$

$$= (10 - 5)^2 \div \left[\frac{20 + 54}{11 - 2} \cdot \frac{11}{5 \times 6} \right] = \frac{25}{3.0148} = 8.29$$

Standard error of the difference.—Until recently, it should be noted here, this type of problem was generally approached in a quite different manner. It has been customary in such problems to calculate the so-called standard error of the difference of two means (σ_D), by which is meant the estimated standard deviation of the differences between numerous means found by repeated drawings of paired samples (see page 170). In practice, this standard deviation of the difference was estimated as

$$\sigma_D = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2}$$

The actual difference, $M_2 - M_1$, was then compared with this σ_D, and if it exceeded 3 times σ_D, the two parent populations were regarded as probably significantly different. Later, the ratio was appraised by reference to a table of t, as has been explained in Chapter VII, where

$$t = \frac{|M_2 - M_1|}{\sigma_D}$$

[1] In terms of the correlation of X and Y depicted in Fig. 19·1, this formula depends upon the following equivalents: $M_2 - M_1 = b$; $\Sigma y_1^2 + \Sigma y_2^2 = \Sigma d^2$; $\Sigma t^2 = b \Sigma xy$; and $N_1 N_2 / N = \Sigma x^2$, as indicated below. Hence it becomes:

$$F = b^2 \div \left[\frac{\Sigma d^2}{N - 2} \cdot \frac{1}{\Sigma x^2} \right] = b \times \frac{\Sigma xy}{\Sigma x^2} \times \frac{N - 2}{\Sigma d^2} \times \Sigma x^2$$

$$= \frac{\Sigma t^2}{\Sigma d^2} (N - 2) = \frac{\Sigma t^2 / \Sigma y^2}{\Sigma d^2 / \Sigma y^2} (N - 2) = \frac{r^2}{1 - r^2} (N - 2)$$

With the X scale as 0 and 1, $\Sigma X = \Sigma X^2 = N_2$ and

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = \frac{N N_2 - N_2^2}{N}$$

Since $N = N_1 + N_2$ this reduces to $N_1 N_2 / N$, or, if $N_1 = N_2$, to $N/4$. Σx^2 is unchanged with other unit-spaced values of X,

with $N - 2$ degrees of freedom, where N is the combined number of items in the two samples.

It may be shown that theoretically this method does not coincide with the correlation method just described, or in general with the analysis of variance except when $N_1 = N_2$, although in other cases it serves as a convenient approximation. Its inexactness lies in the fact that, broadly speaking, it does not utilize the best estimate available of the standard deviation of the universe (σ_u) from which, on the basis of what is called the null hypothesis (see page 171), the two samples are assumed to be drawn.[1] It is for this reason that the older approach has recently been superseded by the analysis of variance method.

Procedure in analysis of variance.—When the problem is set up for analysis of variance, it takes a form such as is presented in Example 19·1. The procedure illustrated is expanded beyond the minimum requirements of computation, in order to facilitate exposition. It may be described as follows: Under each column of data, Y_1 and Y_2, are recorded necessary items or calculations derived from each column separately, and these results are summed in a third column headed "Sums of rows." The first item recorded is N_c, the second ΣY_c, and the third ΣY_c^2, the last being obtained by squaring each item of the data and adding. The next item is the usual correction term

---

[1] When two samples assumed to be drawn from the same universe are available, the best estimate of σ_u is obtained by the formula

$$\sigma_u^2 = \frac{\Sigma y_1^2 + \Sigma y_2^2}{(N_1 - 1) + (N_2 - 1)} = \frac{\Sigma d^2}{N - 2}$$

which is an average of the two estimates utilized in the traditional method, weighted by the degrees of freedom. Then

$$\sigma_{M_1}^2 = \frac{\Sigma d^2}{N_1(N - 2)} \; ; \; \sigma_{M_2}^2 = \frac{\Sigma d^2}{N_2(N - 2)}$$

$$\sigma_D^2 = \sigma_{M_1}^2 + \sigma_{M_2}^2 = \frac{N_2 \Sigma d^2 + N_1 \Sigma d^2}{N_1 N_2(N - 2)} = \frac{N \Sigma d^2}{N_1 N_2(N - 2)}$$

and F (or t^2) is

$$F = (M_2 - M_1)^2 \div \frac{N \Sigma d^2}{N_1 N_2(N - 2)}$$

which is identical with the algebraic simplification of the correlation method (cf. page 458), and also equivalent to the analysis-of-variance method, next to be described.

$(\Sigma Y_c)^2/N_c$, and the difference between the last two items is then recorded as Σy_c^2, meaning the centered squares for each column.

<div align="center">EXAMPLE 19·1</div>

<div align="center">ANALYSIS OF VARIANCE: SIGNIFICANCE OF DIFFERENCE OF
TWO MEANS, UNCORRELATED DATA</div>

Data: Assumed productivity in units made by comparable machine operators in two plants, Y_1 and Y_2 (see Table 19·1).

	Y_1	Y_2		
	4	9		
	2	5		
	5	8		
	8	12	Sums	
	6	14	of	
		12	rows	
N_c	5	6	11	
ΣY_c	25	60	85	$85^2 \div 11 = 656.818 = C$
ΣY_c^2	145	654	799	$- 656.818 = 142.182 = \Sigma y^2$
$(\Sigma Y_c)^2/N_c$	125	600	725	$- 656.818 = 68.182 = \Sigma t^2$
Σy_c^2	20	54	74	$= \Sigma d^2$

<div align="center">MEAN SQUARES</div>

Whole table: $\Sigma y^2/(N-1) = 142.182 \div 10 = 14.218$

Between columns: $\Sigma t^2/(m-1) = 68.182 \div 1 = 68.182$

Within columns: $\Sigma d^2/(N-m) = 74.000 \div 9 = 8.222$

$$F = 68.182 \div 8.222 = 8.29 \text{ (table: 5.12; 10.56)}$$

Check, by revised traditional method:

$$\sigma_D^2 = \frac{\Sigma y_1^2 + \Sigma y_2^2}{N-2} \cdot \frac{N}{N_1 N_2} = \frac{20 + 54}{9} \cdot \frac{11}{5 \times 6} = 3.0148$$

$$F = (M_2 - M_1)^2 \div \sigma_D^2 = (10 - 5)^2 \div 3.0148 = 8.29$$

Up to this point the procedure might suggest that the objective was merely to obtain the mean and standard deviation of the two columns of data separately, as in the traditional method,

inasmuch as the mean would be the second item divided by the first and the standard deviation would be the square root of the last item divided by N. The real purpose, however, is somewhat different. It will be noted that a correction term for the table as a whole is obtained as $C = 85^2/11 = 656.818$. This correction term is subtracted from the aggregate ΣY^2 and it is also subtracted from the sum of the correction terms in the individual columns, which is necessarily greater than the correction term of the whole table. As a result of these two subtractions, the following items are obtained:

$$\Sigma y^2 = 142.182$$
$$\Sigma t^2 = 68.182$$

It is these items, as well as the method of their computation, that require explanation.

If, as has already been suggested, the problem is thought of as a simple correlation of Y_1 and Y_2 combined against X ordinates 0 and 1, or 1 and 2, respectively, the significance of Σy^2 and Σt^2 may readily be seen. In the first place, Σy^2 is merely the aggregate variability of the data as in any problem of simple correlation. Also the correction terms in columns 1 and 2, namely, 125 and 600, respectively, may be thought of as the squared trend points (M_1^2 and M_2^2) on the regression line, multiplied by the frequencies on each ordinate. That is, in the first group $(\Sigma Y_1)^2/N_1 = N_1 M_1^2$, and in the second column $(\Sigma Y_2)^2/N_2 = N_2 M_2^2$. The sum of these two groups of squares is therefore ΣT^2. And ΣT^2, centered, is (since $\Sigma Y = \Sigma T$)

$$\Sigma t^2 = \Sigma T^2 - \frac{(\Sigma Y)^2}{N}$$

By reference to the regression line passing through the means of the data (Fig. 19·1), it may be seen that Σy_1^2 and Σy_2^2 together make up Σd^2 of simple correlation, though usually the deviations are scattered along the regression line rather than grouped on only two ordinates. From the method of calculation as well as from the theory of correlation it may be deduced that the three sums of squares should be related, that is, $\Sigma y^2 =$

$\Sigma t^2 + \Sigma d^2$. Hence Σd^2 as obtained from $\Sigma y_1^2 + \Sigma y_2^2$ should check with $\Sigma y^2 - \Sigma t^2$, or this subtraction alone may be relied upon.

It would now be possible to calculate r as the square root of $\Sigma t^2/\Sigma y^2$, and the significance of the r thus obtained would represent the significance of the difference between the two means. But, for purposes of handling more complex analysis later, it may be better to introduce the measure of significance in a somewhat different form. This form expresses or calculates mean squares, from the three sums of squares in question, by dividing each by its degrees of freedom. In other words, the mean squares regarded as corrected for sampling are calculated.

Degrees of freedom.—The determination of the degrees of freedom, that is, the corrected N for each group of squares, is often a difficult problem. In respect to the whole table, however, it will readily be seen that the degrees of freedom are $N - 1$, since this is an ordinary problem in standard deviation. The degrees of freedom of the third or residual mean square may also be deduced from the fact that the deviations in this case are from a regression line described by two constants. Hence the degrees of freedom are $N - m$, where m means the number of series entering into a correlation ratio, or the number of constants in the regression equation, which in this case is $T = a + bX$. In this instance, therefore, $m = 2$.

The degrees of freedom for the remaining mean square may be inferred from the fact that in effect the group is the unit, and the divisor is therefore $2 - 1 = 1$. Or it may be obtained by noting the general principle that the same breakdown holds for degrees of freedom as for the sums of the squares. In other words, just as $\Sigma y^2 = \Sigma t^2 + \Sigma d^2$, so also the degrees of freedom of Σy^2 must equal the sum of the other two degrees of freedom. Hence the degrees of freedom for regression between columns are $(N - 1) - (N - m) = m - 1$, or in the case at hand, 1.

In practice it is usually unnecessary to calculate the mean square for the whole table, although it is well to record both the sums of the squares and the degrees of freedom, since these are commonly required in determining or checking one of the other sums of the squares or degrees of freedom. It is necessary,

however, to calculate the mean square between columns and within columns, as well as the ratio of these two mean squares, for this ratio is the value of the statistic F. It will be seen that this ratio is merely a comparison of the variance of the groups, as represented by the means, with the best estimate of the variance of the universe from which, on the basis of the null hypothesis, they are assumed to be drawn. That is,

$$F = 68.182 \div 8.222 = 8.29$$

A reading of the table of F in the first column ($m - 1 = 1$) and for the ninth row ($N - m = 9$) shows that the value of F thus obtained is between the least significant and least highly significant levels. Hence the conclusion may be drawn that the level of performance in the second case (Y_2) is significantly above that of the first case (Y_1), though not decisively so.[1] The problem obviously does not isolate the cause of this difference, but merely indicates that a significant difference is present.

Analysis of several means.—An important advantage of analysis of variance as compared with more elementary methods of comparing two means is that it may be extended to apply to a number of different means, or to numerous groups of comparable data which are to be tested for homogeneity. For example, suppose that the central management of a corporation wishes to inquire into the relative efficiency of a certain type of machine operation as carried on in different factories. The personnel department prescribes sample tests and tabulates the resulting scores by factories. It is assumed that the data of Example 19·2 represent such scores in simplified form, that is, the scores listed under Y_1 are from the first factory, under Y_2 from the second factory, and so on. The problem, then, is to determine the significance of the differences among several

---

[1] That F, computed as the ratio of the two mean squares, $\Sigma t^2/(m - 1)$ and $\Sigma d^2/(N - m)$, is the same as F previously computed from r may readily be proved by dividing each term of the ratio by Σy^2, and substituting $\Sigma y^2 - \Sigma t^2$ for Σd^2, as follows:

$$F = \frac{\Sigma t^2}{\Sigma d^2} \times \frac{N - m}{m - 1} = \frac{\Sigma t^2/\Sigma y^2}{\Sigma d^2/\Sigma y^2} \times \frac{N - m}{m - 1}$$

$$= \frac{\Sigma t^2/\Sigma y^2}{(\Sigma y^2 - \Sigma t^2)/\Sigma y^2} \times \frac{N - m}{m - 1} = \frac{r^2}{1 - r^2} \times \frac{N - m}{m - 1}$$

means, that is, whether the factories represent distinctly different levels of performance. The objective, of course, would be the improvement of the backward factories, if it is determined that the differences are somewhat more than the minor variations which might accidentally occur.

EXAMPLE 19·2

ANALYSIS OF VARIANCE: SIGNIFICANCE OF DIFFERENCES AMONG SEVERAL MEANS, UNCORRELATED DATA

Data: Assumed productivity in units of comparable machine operators, groups Y_1, Y_2, Y_3, and Y_4.

	Y_1	Y_2	Y_3	Y_4		
	4	9	9	12		
	2	5	20	5		
	5	8	18	15		
	8	12	14	9		
	6	14	18	16		
		12	11	10		
				17	Row	Correc-
					totals	tions
N_c	5	6	6	7	24	
ΣY_c	25	60	90	84	259	$259^2/24 = 2{,}795.04 = C$
ΣY_c^2	145	654	1,446	1,120	$3{,}365 - 2{,}795.04 =$	$569.96 = \Sigma y^2$
$(\Sigma Y_c)^2/N_c$	125	600	1,350	1,008	$3{,}083 - 2{,}795.04 =$	$287.96 = \Sigma t^2$
Σy_c^2	20	54	96	112	$282 = \Sigma d^2$	

MEAN SQUARES

Whole table: $\Sigma y^2/(N-1) = 569.96 \div 23$

Between columns: $\Sigma t^2/(m-1) = 287.96 \div 3 = 95.987$

Within columns: $\Sigma d^2/(N-m) = 282.00 \div 20 = 14.1$

$$F = 95.987 \div 14.1 = 6.81 \quad (5\% = 3.10; \ 1\% = 4.94)$$

The problem calls for little comment, inasmuch as the procedure is practically the same as in Example 19·1. As before, calculations are made for each factory as if its mean score and standard deviation were to be found. The first row below the tabulated scores lists the number of items, the second row lists

the sum of the items, and the third row the sum of the squares. These squares are then corrected both by groups and as an aggregate. The fourth row gives the individual correction terms, and its total (ΣT^2), corrected, is the sum of the centered squares by groups. That is, Σt^2 is the sum of the centered squares in the whole table as they would be if each worker made the average score in his factory. The fifth row gives the centered squares within each factory, and its sum is obviously Σd^2. It should be noted that $\Sigma d^2 = 282$ should check as the difference of the two row totals directly above.

The table of mean squares follows the same procedure as in Example 19·1. In this case, however, m, the number of columns or groups of workers, is 4, while N, the total number of workers tested, is 24. The mean squares are corrected for sampling as before, and the ratio of the variance between columns to that within columns is $F = 6.81$. As in Example 19·1, this ratio is a comparison of the variance of the groups with a best estimate of the variance of the universe from which they are assumed by the null hypothesis to be drawn. As indicated by the degrees of freedom, the table of F is read at column 3, row 20, where the least significant and least highly significant values are given as 3.10 and 4.94, respectively. Hence, it may be concluded that the performance of workers in different factories is at significantly different levels.

An inspection of the procedure in Example 19·2 will show that, like Example 19·1, where only two groups were compared, this also may be described as a problem in determining the significance of correlation. But in this case the correlation is so complex that the comparison does not greatly clarify the explanation. However, the process may be described as involving the correlation ratio, or, if the X scale is taken as 0, 1, 2, etc., it may be regarded as a curvilinear correlation with a potential trend line of the third order passing through the mean of each column. The statistic F therefore measures the reliability with which the regression line describes the distribution of the data. That is, Σy^2 is the sum of the centered squares in the whole table; Σt^2 is the sum of the centered squares of the regression values; and Σd^2 is the sum of the alienation squares $(Y - T)^2$.

Hence, the three sums of centered squares represent typically the entire table, the regression (or groups), and the deviations from that regression, respectively.

However, it is probably more satisfactory to drop the analogy to correlation at this point. It is necessary only to note that the analysis of variance answers the question whether the groups as such differ among themselves more than would be expected from the composite variabilities of numerous individual groups. A significant value of F indicates that on the basis of chance drawings from the same universe such divergences among the groups would rarely occur.

Analysis applied to grouped data.—In Example 19·3, the same type of problem is again presented, the only difference being that the data have been grouped. That is, the two sets of frequencies, f_1 and f_2, are two ordinary frequency distributions with the class marks shown as Y. The process of calculation is as before except as the arrangement of the data determines minor changes.

Below the tabulated data, the successive rows record the same items as before, namely, the number of items in each series, N_c, the sums of the items, ΣY_c, and the sums of the squares, ΣY_c^2. The fourth row lists the correction item for each of the two sums of squares (it is $\overline{\Sigma Y_c^2}/N_c$), and the fifth row is the sums of the centered squares.

An additional column included with the data makes available a new numerical check, namely, the total frequencies at each class mark. These frequencies may be used to check the first three of the items recorded below them, which are also the row sums. For example, the aggregate N is the sum of both the column and the row in which it appears. The aggregate Y's, 552, is both the sum of $232 + 320$ and the sum of the total frequencies (by rows) times Y. The Y^2 aggregate also checks in the same way.

Next, the aggregate Y squares in the third row of the calculations are reduced by subtracting their correction term, $(\Sigma Y)^2/N = 6771.2$, where both Y and N refer to the whole table. The result is the centered squares, representing the deviations of all the items from the mean of the whole table.

The fourth row sums the two group correction terms and reduces the total by the same correction term as before. The result may be described as the sum of centered squared deviations

EXAMPLE 19·3

ANALYSIS OF VARIANCE: SIGNIFICANCE OF DIFFERENCE OF
TWO MEANS, UNCORRELATED GROUPED DATA

Data: Assumed productivity in units (Y) made by comparable machine operators, groups indicated by frequencies f_1 and f_2.

Y	f_1	f_2	Total f
10	3	1	4
11	7	3	10
12	6	5	11
13	3	8	11
14	1	7	8
15	0	1	1
N_c	20	25	45
ΣY_c	232	320	552
ΣY_c^2	2714	4132	$6846.0 - 6771.2 = 74.8 = \Sigma y^2$
$(\Sigma Y_c)^2/N_c$	2691.2	4096	$6787.2 - 6771.2 = 16.0 = \Sigma t^2$
Σy_c^2	22.8	36	$58.8 = \Sigma d^2$

MEAN SQUARES

Whole table: $\Sigma y^2 \div (N - 1) = 74.8 \div 44$
Between columns: $\Sigma t^2 \div (m - 1) = 16.0 \div 1 = 16.0$
Within columns: $\Sigma d^2 \div (N - m) = 58.8 \div 43 = 1.3674$

$$F = 16 \div 1.3674 = 11.70 \text{ (table: 4.06; 7.23)}$$

Check:

$$\sigma_D^2 = \frac{\Sigma y_1^2 + \Sigma y_2^2}{N - 2} \cdot \frac{N}{N_1 N_2} = \frac{22.8 + 36.0}{43} \cdot \frac{45}{20 \times 25} = 0.1231$$

$$F = (M_2 - M_1)^2 \div \sigma_D^2 = (12.8 - 11.6)^2 \div 0.1231 = 11.70$$

with each individual scored at the average of his group. It thus represents the variability of the groups as such, the variability within the groups being eliminated. The fifth row lists

the centered squares from the group means, aggregating 58.8, which may be checked as the difference of the two items immediately above.

The calculation of the mean squares does not require detailed description. The degrees of freedom are calculated as explained, and the sum of the second two equals the first. The statistic F is calculated as the mean square measuring group variability compared with the mean square measuring individual variability. As in the preceding problem, the value of F is compared with the result found by use of the formula for σ_D^2. By reference to the table of F it is found that the difference of the group means is highly significant.

Analysis applied to several groups.—As has previously been noted, it is one of the advantages of the analysis of variance that it may be applied to comparisons among a number of different groups as well as between two groups. In other words, it answers the question whether there are significant differences among many means as well as between two means. This application has many practical uses, as when performance or status in various groups or numerous geographic areas is to be compared. The process as illustrated in Example 19·4, however, so closely parallels that of 19·3 that it does not require much additional description. It will be seen that the five rows of calculations following the tabulated data are arranged exactly as before, and the corrections in the row totals also follow the earlier procedure. In the table of mean squares, however, it may be observed that the degrees of freedom reflect the increase in the number of groups. For the whole table they are $N-1$, as before. But among columns, where in effect the group is the unit, they are $m - 1 = 4 - 1 = 3$. And for the variability within columns they are found either as the difference between the two preceding degrees of freedom $(64 - 3 = 61)$ or as $N - m = 65 - 4 = 61$.

If F is computed as a measure of the relative group variability, as before, a value of 0.64 results. This fractional value suggests that group variability is less than might be expected by chance. In such a case it is customary to recalculate F as the reciprocal of its discovered value (i.e., $F = MS_d \div MS_t =$

1.57), in accordance with its definition as the ratio of greater to lesser constituent mean square. As such, it is shifted to another sampling distribution, the significant limiting values of which are read from the table by a direct interchange of column and row. For the case at hand, therefore, the least

<div align="center">Example 19·4</div>

ANALYSIS OF VARIANCE: SIGNIFICANCE OF THE DIFFERENCES AMONG SEVERAL MEANS, UNCORRELATED GROUPED DATA

Data: Assumed productivity in units, coded as Y, made by comparable machine operators classified in 4 (m) groups, f_1, f_2, f_3, and f_4.

Score	Y	f_1	f_2	f_3	f_4	Σf	
10	0	0	1	0	0	1	
11	1	3	3	2	1	9	
12	2	7	5	4	5	21	
13	3	6	8	3	3	20	
14	4	3	7	1	1	12	
15	5	1	1	0	0	2	
N_c		20	25	10	10	65	
ΣY_c		52	70	23	24	169	
ΣY_c^2		158	232	61	64	515	$-439.4 = 75.6 = \Sigma y^2$
$(\Sigma Y_c)^2/N_c$		135.2	196.0	52.9	57.6	441.7	$-439.4 = 2.3 = \Sigma t^2$
Σy_c^2		22.8	36.0	8.1	6.4	73.3	$= \Sigma d^2$

<div align="center">Mean Squares</div>

Whole table: $\Sigma y^2/(N-1) = 75.6 \div 64$
Between columns: $\Sigma t^2/(m-1) = 2.3 \div 3 = 0.7667 = MS_t$
Within columns: $\Sigma d^2/(N-m) = 73.3 \div 61 = 1.2016 = MS_d$

Group variability: $F = 0.7667 \div 1.2016 = 0.64$ (table: 2.76; 4.13)
Group similarity: $F = 1.2016 \div 0.7667 = 1.57$ (table: 8.6; 26.3)

significant value (col. 61, row 3) is between 8.58 and 8.53, and the discovered measure is distinctly not significant. The two measures of F thus obtained indicate (1) that variability among the groups is not significantly greater than what might be expected by chance and (2) that their similarity or lack of variability is also not significant.

The interpretation obviously is that the levels of performance in the various factories, taken as a whole, do not vary significantly. If, however, $F = MS_d \div MS_t$ had been highly significant, this would have meant that the means were more alike than would be expected by chance. Under some such circumstances a "doctoring" of the figures might be suspected.

Analysis of correlated data.—It has been noted that the methods just described apply to two sets of data between which no correlation is logically to be expected or to several sets which are comparable but not correlated. It was also stated that adjustments are necessary if such analysis is to be applied to two or more correlated series. This problem may be illustrated by reference to two series whose items are paired. Such series are illustrated in Example 19·5 where a short-cut equivalent of analysis of variance is applied to the records of a certain group of machine operators on two different dates. The first set of scores (Y_1) represents the productivity of the workers before taking a course of training, and the second set of scores (Y_2) measures their efficiency after the training. The pairing of the items arises from the fact that each row is made up of two records for the same workman. The problem obviously is to determine whether the training significantly increased efficiency, that is, whether the mean of the second group is greater than the mean of the first group to a degree not accounted for by mere chance variability. Except for the existence of correlation, the problem is equivalent to Example 19·1.

The approach to the problem modifies the analysis of variance technique in that the gains (positive or negative) made by each workman, presumably as a result of training, are utilized in discovering the value of F. In ·fact, F is here merely the squared mean gain divided by its own variance, the squared standard error of the mean. The logic is clearer, however, if t rather than F $(t = \sqrt{F})$ is employed. Assuming a null hypothesis, the difference of the means $(M_2 - M_1)$, or, what is the same thing, the mean of the individual gains (M_g), is expected to be zero. Its sampling variability is the standard error of that mean (σ_{M_g}). If this standard error were accurately known, if it could be measured instead of estimated, the ratio $M_g \div \sigma_{M_g}$

could be evaluated in terms of the probability of deviation (x/σ) in a normal distribution. But since it is estimated from a sample, t should be used, and the critical values of t read from the table at row $N - 1$, where N is the number of gains (not scores). The statistic F is, of course, merely the square of t; that is, $F = M_g^2 \div \sigma_{M_g}^2$.

EXAMPLE 19·5

DUAL ANALYSIS OF VARIANCE: SIGNIFICANCE OF DIFFERENCE BETWEEN THE MEANS, CORRELATED DATA. SHORT-CUT METHOD

Data: Assumed scores made by a group of machine operators (A, B, C, etc.) before a course of training (Y_1), and after the training (Y_2).

Operator	Y_1	Y_2	Gain $(\dot{G})$	G^2
A	3	9	6	36
B	6	12	6	36
C	7	11	4	16
D	9	15	6	36
E	15	13	−2	4

$$\begin{array}{cccc} 5)\overline{40} & 5)\overline{60} & 5)\overline{20} & 128 \\ M_1 = 8 & M_2 = 12 & M_g = 4 & C = 80 \end{array}$$

$$5)\overline{48} = \Sigma g^2$$
$$4)\overline{9.6} = \sigma_g^2$$
$$2.4 = \sigma_{M_g}^2$$

$$F = M_g^2 \div \sigma_{M_g}^2 = 4^2 \div 2.4 = 6.67$$

Degrees of freedom, $N_g - 1 = 4$

NOTE: The above method is an abbreviation of the regular analysis of variance method (see Example 19·7), and is algebraically equivalent to the traditional method. as follows:

$$\sigma_D^2 = \sigma_{M_1}^2 + \sigma_{M_2}^2 - 2r\sigma_{M_1}\sigma_{M_2} = 4 + 1 - (2 \times 0.65 \times 2 \times 1) = 2.4$$

$$F = (M_2 - M_1)^2 \div \sigma_D^2 = 4^2 \div 2.4 = 6.67$$

It may be noted that the traditional approach to the foregoing problem calculates the standard error of the difference of the means, thus:

$$\sigma_D^2 = \sigma_{M_1}^2 + \sigma_{M_2}^2 - 2r_{12}\sigma_{M_1}\sigma_{M_2}$$

which may be shown algebraically to be equivalent [1] to $\sigma^2_{M_g}$. From the standpoint of ease of calculation, however, the latter method is superior. Nevertheless, the traditional formula serves to stress the importance of the degree of correlation. Obviously, the significance of the difference of the means depends not only on the size of the samples and the lack of variability in the items but also on the degree of consistency of the gains as measured by r. The calculation of F is merely a device for summing up these aspects of the evidence. In the method of gains, the lack of correlation (r_{12}) registers in σ_{M_g}.

Dual variance.—The commonly used analysis of variance method, which is applicable to two or more correlated series of data, is illustrated in Example 19·6. The data represent production records for the same group of workmen at different times, as in early and late morning and early and late afternoon. The question is whether the results vary significantly.

The special feature of the method lies in the application of the analysis to both columns and rows. Computation begins with a measure of the variability by columns, precisely as in earlier examples (cf. Example 19·2, page 464), except that the first and last rows, N and Σy^2, are omitted. The total squares in the table (ΣY^2) are 3,452, which, less the correction term, $(\Sigma Y)^2/N = 3,380$, is the centered squares of the table (72).

---

[1] The identity of the two methods of finding σ_D may be proved as follows (taking $N_1 = N_2 = N$, and the two series as X and Y):

A. By traditional method:

$$\sigma_D^2 = \sigma_{Mx}^2 + \sigma_{My}^2 - 2r_{xy}\sigma_{Mx}\sigma_{My}$$

$$= \frac{\Sigma x^2 + \Sigma y^2}{N(N-1)} - \frac{2\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \cdot \frac{\sqrt{\Sigma x^2 \Sigma y^2}}{N(N-1)} = \frac{\Sigma(x-y)^2}{N(N-1)}$$

B. By standard error of mean gain:

$$\sigma_D^2 = \frac{\Sigma G^2 - (\Sigma G)^2/N}{N(N-1)} = \frac{\Sigma(Y-X)^2 - (\Sigma Y - \Sigma X)^2/N}{N(N-1)}$$

Expanded, the second term of the numerator provides the correction terms for the first:

$$\sigma_D^2 = \frac{(\Sigma Y^2 - (\Sigma Y)^2/N) - (2\Sigma XY - 2\Sigma X\Sigma Y/N) + (\Sigma X^2 - (\Sigma X)^2/N)}{N(N-1)}$$

$$= \frac{\Sigma(x-y)^2}{N(N-1)}$$

EXAMPLE 19·6

DUAL ANALYSIS OF VARIANCE: SIGNIFICANCE OF DIFFERENCES AMONG SEVERAL MEANS, CORRELATED DATA

Data: Assumed records made by comparable machine operators (A, B, C, etc.) at different times of day. Successive tests are indicated by Y_1, Y_2, Y_3, and Y_4.

A. Ordinary method:

Operators	Y_1	Y_2	Y_3	Y_4	ΣY_r	ΣY_r^2	$(\Sigma Y_r)^2/N_r$
A	11	12	13	12	48	578	576
B	12	14	16	14	56	792	784
C	14	17	16	13	60	910	900
D	12	15	13	12	52	682	676
E	11	12	12	9	44	490	484
ΣY_c	60	70	70	60	260	3,452	$3,420 - 3,380 = 40 = \Sigma t_r^2$
ΣY_c^2	726	998	994	734	$3,452 - 3,380 = 72 = \Sigma y^2$		
$(\Sigma Y_c)^2/N_c$	720	980	980	720	$3,400 - 3,380 = 20 = \Sigma t_c^2$		

B. Abbreviated method:

Operators	Y_1	Y_2	Y_3	Y_4	ΣY_r	$(\Sigma Y_r)^2$	Whole table:
A	11	12	13	12	48	2,304	$\Sigma Y^2 = 3,452$
B	12	14	16	14	56	3,136	$C = 3,380$
C	14	17	16	13	60	3,600	$\Sigma y^2 = 72$
D	12	15	13	12	52	2,704	
E	11	12	12	9	44	1,936	
ΣY_c	60	70	70	60	260	$13,680 \div 4 = 3,420$; less 3,380 $= 40 = \Sigma t_r^2$	
$(\Sigma Y_c)^2$	3,600	4,900	4,900	3,600	$17,000 \div 5 = 3,400$; less $3,380 = 20 = \Sigma t_c^2$		

MEAN SQUARES

Whole table: $\Sigma y^2/(N-1) = 72 \div 19$

Between columns: $\Sigma t_c^2/(m_c - 1) = 20 \div 3 = 6.67$

Between rows: $\Sigma t_r^2/(m_r - 1) = 40 \div 4 = 10.00$

Residuals: $12 \div 12 = 1.00$

    By columns: $F_{c.r} = 6.67 \div 1 = 6.67$ (table: 3.49; 5.95)

    By rows: $F_{r.c} = 10 \div 1 = 10$ (table: 3.26; 5.41)

This calculation is independent of the arrangement of the data in columns and rows. The sum of the correction terms of the individual columns is 3,400, which, similarly corrected, is the total of squares between columns ($\Sigma t_c^2 = 20$). The variability within groups (Σy_1^2, Σy_2^2, etc.) is not found because the sum of residual squares (Σd^2) of dual variance is modified by the variability of both the groups and the rows (the individual workmen), next to be measured.[1]

The variability by rows is found by the method just applied to columns. The first two totals, summing ΣY_r and ΣY_r^2, should obviously check with the first two totals derived from the columns, summing ΣY_c and ΣY_c^2. The centered squares between rows is then obtained by correcting the aggregate row correction terms. These squares, $\Sigma t_r^2 = 40$, measure the variability of the workers, each measured by the average of his tests.

In section B, still further abbreviation is illustrated. The squares of the whole table are taken care of separately at the right. The aggregate individual correction term by columns, $\Sigma T_c^2 = \Sigma(\Sigma Y_c)^2/N_c$, is found by first summing the squares of the column totals $\Sigma(\Sigma Y_c)^2$, and then dividing this sum by the common divisor, N_c, thus reducing the number of computations. The row corrections are similarly treated. Otherwise, the method is the same as before.

In calculating F, mean squares for the table, for the variability of columns and of rows, and for residual variability, are obtained. The first three measures have already been described; the fourth is merely the part of the total variability of the table unaccounted for by column and row variability, that is $\Sigma d^2 = \Sigma y^2 - \Sigma t_c^2 - \Sigma t_r^2$. The degrees of freedom correspond to those of simple variance; for the table, $N - 1$; for the columns, $m_c - 1$ (the number of columns less 1); for the rows, $m_r - 1$ (the number of rows less 1); and for the residual the balance not

---

[1] It is important to note that dual residual variability (Σd^2) will *not* be the sum of the residual variabilities which might have been obtained by columns and rows. The reason for this discrepancy is that the column and row variabilities interact to explain the actual variability of the table. A case might theoretically arise where dual variability gives no residual variation, but where the variability by columns and rows taken separately might be very considerable. This will become more evident as the nature of regression in dual analysis of variance is described.

required for columns and rows, that is, $(N - 1) - (m_c - 1) - (m_r - 1) = N - m_c - m_r + 1$. The number of residual degrees of freedom may be checked as $(m_c - 1)(m_r - 1)$. The value of F representing variability by columns—that is, the variability of mean efficiency at different times—is found as the ratio of the mean square by columns to the residual mean square $(F_{c \cdot r} = MS_{tc} \div MS_d)$. The critical values are read from the the table, as before, by reference to the degrees of freedom involved (col. 3, row 12: 3.49; 5.95). The significance of the variability of the workers $(F_{r \cdot c})$ may be similarly tested.

The symbol $F_{c \cdot r}$ suggests an analogy to partial correlation, and in fact such an analogy is justified. In effect, dual variance between columns is computed after variability between rows has been eliminated. This is shown by the fact that, if the latter variability is eliminated by reducing each row to the average row, Σt_r is reduced to zero, and Σy^2 is diminished by the same amount.[1]

Similarity to correlation.—It may be worth while, as a means of explaining the nature of dual analysis of variance, to regard it tentatively as a form of correlation. Considered as a whole, without elimination of row or column variability, dual variance is a comparison of the data with a regression pattern (T or Y') described as

$$Y' = M_c + M_r - M_y$$

which means that for any given item (Y) the regression item (Y') is the sum of the mean of the column and the mean of the row in which the item is located, less the mean of the table. Or, if the data are centered (measured as deviations from their mean), the regression item of any y is merely the sum of its column and row means, that is,

$$Y' - M_y = (M_c - M_y) + (M_r - M_y)$$

---

[1] It may easily be verified that, if Example 19·5 is solved by the method of Example 19·6, $\quad \Sigma y^2 = 140; \ \Sigma t_c^2 = 40; \ \Sigma t_r^2 = 76; \ \Sigma d^2 = 24$

And if row variability is eliminated from the data, the columns Y_1 and Y_2, and the sums of squares, become (each paired sum, 20; each difference, G)

$$Y_1 = \ 7, \ \ 7, \ \ 8, \ \ 7, 11$$
$$Y_2 = 13, 13, 12, 13, \ 9$$
$$\Sigma y^2 = 64; \ \Sigma t_c^2 = 40; \ \Sigma t_r^2 = 0; \ \Sigma d^2 = 24$$

The regression values of Example 19·6, and the departures of the data from regression (alienation, $d = Y - Y'$), are listed in Example 19·7. For purposes of comparison the aggregate

<div align="center">

EXAMPLE 19·7

REGRESSION $(Y' = M_c + M_r - M_y)$ AND ALIENATION $(d = Y - Y')$
IN DUAL ANALYSIS OF VARIANCE AND CORRELATION
MEASURES

</div>

Data: See Example 19·6.

DATA (Y)				REGRESSION (Y')				ALIENATION $(Y - Y')$			
Y_1	Y_2	Y_3	Y_4	Y_1'	Y_2'	Y_3'	Y_4'	d_1	d_2	d_3	d_4
11	12	13	12	11	13	13	11	0	−1	0	1
12	14	16	14	13	15	15	13	−1	−1	1	1
14	17	16	13	14	16	16	14	0	1	0	−1
12	15	13	12	12	14	14	12	0	1	−1	0
11	12	12	9	10	12	12	10	1	0	0	−1

$\Sigma Y = 260$; $\Sigma Y^2 = 3{,}452$ $\Sigma Y' = 260$; $\Sigma(Y')^2 = 3{,}440$ $\Sigma d^2 = 12$

$(\Sigma Y)^2/N = 3{,}380$ $(\Sigma Y)^2/N = 3{,}380$

$\Sigma y^2 = \quad 72$ $\Sigma(y')^2 = \quad 60$

$$\rho^2 = \frac{\Sigma(y')^2}{\Sigma y^2} = \frac{60}{72} = 0.8333; \quad F = \frac{\rho^2}{1 - \rho^2} \times \frac{N - m}{m - 1} = \frac{0.8333}{0.1667} \times \frac{12}{7} = 8.57$$

$DF = (m_c + m_r - 2)$ and $(m_c - 1)(m_r - 1) = 7$ and 12. Table: 2.92; 4.65

$$\eta^2_{c.r} = \frac{\Sigma t_c^2}{\Sigma y^2 - \Sigma t_r^2} = \frac{20}{72 - 40} = 0.625;$$

$$F = \frac{\eta^2}{1 - \eta^2} \times (N_c - 1) = \frac{0.625}{0.375} \times 4 = 6.67$$

$DF = (m_c - 1)$ and $(m_c - 1)(m_r - 1) = 3$ and 12. Table: 3.49; 5.95

correlation ratio and limited η on which $F_{c.r}$ is based are also stated.

Analysis of seasonality.—Analysis of variance is a flexible tool, adaptable to a large variety of situations. In business statistics, however, it has not been widely used except in the types of problems already described and in measuring the significance of such patterns as those of seasonality.

A crude approach to a measurement of seasonality may be made by applying dual analysis of variance to data thought to exhibit a seasonal pattern and set up with months or quarters as columns, and years as rows, or vice versa. As an illustration it might be assumed that data of Example 19·6 represent quarterly production data for the years 1934–1938, each row representing a year. The first column then represents the first quarters (January–March), the second column the second quarters (April–June), etc. The significance of column variability as represented by $F_{c.r}$ may therefore serve to indicate whether or not there is a significant seasonal pattern. A pronounced straight-line trend, however, will be reflected in the column variability and may greatly exaggerate $F_{c.r}$, while cyclic variability may unduly diminish it. A straight-line trend might be removed, but more convenient methods are available.

It has been customary to apply the analysis of variance to the seasonal relatives (data divided by annual moving average) rather than to the crude data as suggested above. This procedure tends to eliminate both trend and cyclic variability. If, again, the data of Example 19·6 are assumed to represent seasonal relatives set up with fiscal years in the rows (SR begins and ends at the middle of the original calendar year), it is evident that significant column variability is the measure of seasonal pattern. It has been objected that this method is mathematically inexact because it fails to take into account the loss of squares and degrees of freedom involved in the removal of the moving average, and also because the ratios thus obtained will not be normally distributed and that spurious correlation may be involved.[1] On the other hand, it is argued that the seasonal relatives are indirect measures of the "seasonal thrust" and

[1] The term "spurious correlation" means a degree of covariation that is attributable to the method of securing or manipulating the data rather than to any interrelationship. It frequently reflects the results of multiplying or dividing each pair or series of correlative items by a specific factor. A general factor, the same for all items, would, of course, here be ineffective. No matter how lacking in correlation the original items may be, the specific factors thus applied may produce significant correlation. However, the factor $1/MA$ which appears in the seasonal relatives is not constant for correlative months, as the Januaries. It is a measure of what the time series would have been if seasonal forces had not driven it temporarily from its course.

thus may be regarded as original data. Their distribution also is said to approximate a skewed normal as closely as other data to which analysis of variance is commonly applied. Experience shows that the method in question does in fact register through F the uniformity of the seasonal pattern. Even if the limiting probabilities are not exactly as stated in the table, others could be set up on the basis of experience.

The ranking method.—Fortunately a method adaptable to seasonality, and largely free from the technical difficulties encountered in the method just described, is available. This is the so-called chi-square ranking test. It is a method of dual analysis which eliminates year-to-year variability by the simple expedient of substituting ranks for seasonal relatives within each year. It has already been mentioned (see page 299). It should be noted, however, that the method is generally applicable to problems where departure from normality is an obstacle.

The chi-square ranking test is illustrated in Example 19·8. The items on which the rankings are based are the seasonal relatives of the strike data indicated. For example, the first column of the seasonal relatives, for the fiscal year July, 1927, to June, 1928, together with their rankings, is as follows:

Mo.:	July	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
SR:	93.4	93.7	99.4	89.4	51.2	61.8	89.3	91.1	81.6	139.6	159.2	86.3
Rank:	8	9	10	6	1	2	5	7	3	11	12	4

Other years are ranked in the same way. Ties may generally be resolved by carrying moving averages and seasonal relatives to more decimal places, or, if this is not convenient, they may be ranked in the order in which they occur and the ranks averaged. Thus if the first SR listed above had been 93.7 the first two ranks would be 8.5 each instead of 8 and 9. Obviously the totals of ranks for each year (in Example 19·8, columns are set up as years) are identical, that is, $\Sigma R = R(R + 1)/2 = 12 \times 13/2 = 78$. Hence year-to-year variability—in this case column variability—is eliminated.

The totals of the ranks for each month are then noted. They may be labeled X. It is evident that X variability reflects seasonality, which is merely a tendency toward uniformity in the

<div align="center">EXAMPLE 19·8</div>

SIGNIFICANCE OF SEASONALITY MEASURED BY CHI SQUARE APPLIED TO RANKING OF SEASONAL RELATIVES ($Y \div MA$) WITHIN EACH FISCAL YEAR

Ranks from smallest (1) to largest (12). X = sum of rankings for each month.

Data: Number of strikes beginning each month, January, 1927 through December, 1936 (cf. "Seasonality in Strikes," by Dale Yoder, *Journal of the American Statistical Association*, December, 1938, pages 687–693.

Month	Rank of month in 12-month period ending July 1										
	'28	'29	'30	'31	'32	'33	'34	'35	'36	X	X^2
July	8	9	8	11	6	8	11	6	9	76	5,776
Aug.	9	8	9	7	8	12	12	12	12	89	7,921
Sept.	10	5	12	9	9	11	10	5	4	75	5,625
Oct.	6	10	6	3	5	4	5	11	10	60	3,600
Nov.	1	2	5	2	2	2	2	2	2	20	400
Dec.	2	1	1	1	1	1	1	1	1	10	100
Jan.	5	4	2	6	11	7	3	3	5	46	2,116
Feb.	7	3	3	5	3	3	4	4	3	35	1,225
Mar.	3	6	4	4	4	6	6	8	8	49	2,401
Apr.	11	12	11	10	10	5	8	9	6	82	6,724
May	12	11	7	12	12	10	9	7	11	91	8,281
June	4	7	10	8	7	9	7	10	7	69	4,761

$$702 \qquad 48,930$$
$$(\Sigma X)^2/N_x = \overline{41,067}$$
$$\Sigma x^2 = \overline{7,863}$$

$$\chi_r^2 = \frac{6\Sigma x^2}{\Sigma X} = \frac{6 \times 7,863}{702} = 67.21$$

Table, $n = 11$: 1% probability, $\chi^2 = 25$.

Seasonality may also be measured by correlation ($\chi^2 \div$ maximum χ^2):

If ρ = the number of ranks (months) and n_s = the number of sets of ranks (years)

$$\eta_r^2 = \frac{\chi_r^2}{n_s(\rho - 1)} = \frac{67.21}{9(12 - 1)} = 0.6789$$

$$\eta_r = \sqrt{0.6789} = 0.8240$$

ranking of seasonal relatives in each year. But, partly because ranks do not form a normal distribution, it is found more convenient to measure the significance of its variability by means of chi square rather than F. For a normal distribution the chi-square measure of variability of X from its mean would reduce to $N\Sigma x^2/\Sigma X$. But, because of the nature of the X series, it becomes:

$$\chi_r^2 = \frac{6\Sigma x^2}{\Sigma X}$$

which would agree with the ordinary chi square only if $N = 6$, that is, if 6 seasonal intervals were employed (January–February; March–April; etc.).

The computation of chi square in Example 19·8 needs no explanation. With adequate data (at least a 4×5 table) it may be interpreted conventionally in terms of a chi-square table by reading the 5 and 1 per cent probabilities for $n = N - 1$. In the case at hand these probabilities require approximately 20 and 25. They are the limiting values applicable to all problems of monthly seasonality, if the usual probability standards are to be applied.

READINGS

BRANDT, A. E., "The Analysis of Variance in a '2 $\times$ s' Table with Disproportionate Frequencies," *Journal of the American Statistical Association*, 28 (182), June, 1933, pp. 164–174.

COCHRAN, W. G., "The Use of the Analysis of Variance in Enumeration by Sampling," *Journal of the American Statistical Association*, 34 (207), September, 1939, pp. 492–510.

FRIEDMAN, MILTON, "The Use of Ranks to Avoid the Assumption of Normality," *Journal of the American Statistical Association*, 32 (200), December, 1937, pp. 675–701.

IRWIN, J. O., "Mathematical Theorems Involved in the Analysis of Variance," *Journal of the Royal Statistical Society*, 94 (2), 1931, pp. 284–300.

PEARSON, EGON S., "The Test of Significance for the Correlation Coefficient," *Journal of the American Statistical Association*, 26 (174) June, 1931, pp. 128–134.

PEPPER, JOSEPH, "Studies in the Theory of Sampling," *Biometrica*, 21 (1–4), December, 1929, pp. 231–258.

RIDER, PAUL, "A Survey of the Theory of Small Samples," *Annals of Mathematics*, 31 (4) October, 1930, pp. 577–628.

SCHULTZ, T. W., and SNEDECOR, G. W., "Analysis of Variance as an Effective Method of Handling the Time Element," *Journal of the American Statistical Association*, 28, March, 1933, pp. 14–30.

SNEDECOR, GEORGE W., *Calculation and Interpretation of Analysis of Variance and Covariance*, Ames, Iowa, Collegiate Press, 1934.

WALLIS, W. ALLEN, "The Correlation Ratio for Ranked Data," *Journal of the American Statistical Association*, 34 (207), September, 1939, pp. 533–538.

ZUBIN, JOSEPH, "Nomographs for Determining the Significance of the Differences between the Frequencies of Events in Two Contrasted Series or Groups," *Journal of the American Statistical Association*, 34 (207), September, 1939, pp. 539–544.

EXERCISES AND PROBLEMS

A. EXERCISES

1. Assuming each of the following exercises to represent two series of test scores of comparable workmen, determine whether the means of the series are significantly different.

(a)		(b)		(c)		(d)		(e)		(f)	
Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
1	14	2	10	2	10	1	2	1	2	1	7
3	18	4	12	6	12	3	4	3	8	3	8
	18		12		12		4		10		8
	30		18		18		10		11		10
									14		11
											16

(g)		(h)		(i)		(j)		(k)		(l)	
Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
1	11	1	8	1	6	1	9	1	15	1	18
3	15	3	12	5	10	5	13	5	19	5	22
	15		12		10		13		19		22
	27		24		22		25		31		34

2. For the following sets of uncorrelated data, determine the significance of the difference between the means.

(a)		(b)		(c)		(d)		(e)		(f)	
1	9	2	4	2	6	2	9	2	1	2	5
9	10	6	8	3	8	3	10	3	6	2	7
	11		10	5	9	3	10	6	6	7	10
			12	6	9	4	14	9	12	8	16
					10		15		15	10	20
					12		20		20	13	20

(g)		(h)	
1	3	3	1
5	9	5	3
8	13		5
10	13		8
11	15		10
13	19		10
			15
			20

3. For the following sets of uncorrelated data, determine the significance of the variability of the means.

(a)			(b)			(c)		
Y	f_1	f_2	Y	f_1	f_2	Y	f_1	f_2
0	1	1	0	2		0	2	1
1	2	3	1	4	2	1	4	3
2	1	5	2	3	6	2	2	4
3	1	1	3	1	10	3		3
			4		2	4		1

(d)				(e)				(f)			
Y	f_1	f_2		Y	f_1	f_2		Y	f_1	f_2	f_3
0	3	0		0	0	0		0	1	1	1
1	2	1		1	2	1		1	3	2	3
2	1	4		2	4	3		2	1	1	5
3	1	4		3	2	4		3		1	1
4		2		4		3					
5		2		5		1					

(g)			
Y	f_1	f_2	f_3
1	4	1	3
2	8	5	6
3	6	3	1
4	2	1	

(h)			
Y	f_1	f_2	f_3
0	1		1
1	4	1	2
2	6	1	1
3	4	5	
4	1	3	

(i)			
Y	f_1	f_2	f_3
0	2	1	
1	4	4	
2	3	6	1
3	1	4	3
4		1	5
5			1

(j)			
Y	f_1	f_2	f_3
0	2		
1	4	1	
2	3	4	2
3	1	6	3
4		4	4
5		1	1

(k)

Y	f_1	f_2	f_3	f_4
12				1
9		2	4	1
6	2	3	1	
3	1	1		

(l)

Y	f_1	f_2	f_3	f_4	f_5
5		1	6	1	
4		3	10	5	
3		4	4	3	1
2	1	2		1	4
1	4				

(m)

Y	f_1	f_2	f_3	f_4	f_5
4		1	2	3	
3		2	7	3	1
2	3	4	0	1	2
1	2	2	1	1	1
0	3	1			

4. From factories A, B, C, etc., samples of scores made by a certain class of machine operators were obtained, as indicated below. Do the scores for the individual factories show a significant variability?

Y	A f_1	B f_2	C f_3	D f_4	E f_5	F f_6	G f_7	H f_8	I f_9	J f_{10}	K f_{11}
1	2	1	1	3	2	1	3	3	5	8	9
2	3	4	1	3	4	5	5	6	5	2	1
3	3	3	6	2	3	3	2	1			
4	1	1	1	2	1	1					
5	1	1	1								

5. Analyze the following sets of uncorrelated data to determine the significance of the difference between the means by single variance.

(a)		(b)		(c)		(d)		(e)	
Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
1	5	10	9	2	3	4	2	3	4
3	6	12	11	6	8	3	5	14	18
5	7	8	10	4	9	1	9	3	7
		14	15	3	9	5	11	11	16
		16	20	5	11	7	13	9	15
								8	12

(f)		(g)		(h)		(i)		(j)	
Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
2	4	5	13	9	21	11	26	10	25
6	10	2	6	4	10	10	15	16	15
9	14	6	12	6	11	12	19	12	14
11	14	3	6	9	13	8	31	9	16
12	16	8	13	11	21	9	33	11	28
14	20	6	10	13	22	6	30	5	25
				8	12	11	30	15	20
				4	10	13	16	10	9

(k)		(l)		(m)		(n)	
Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
13	20	15	32	11	16	14	21
6	9	4	24	4	2	6	10
8	10	5	28	6	7	10	11
11	12	7	30	9	24	12	13
13	20	15	32	11	27	14	21
14	21	16	38	12	28	20	22
13	11	6	30	11	21	12	12
2	9	4	26	0	3	8	10

6. Assuming the data of Exercise 5 to be correlated, determine the significance of the variability by dual variance (a) by comparison of the means and (b) by finding the standard error of the mean of the gains.

Answers to Exercises

	F	l.s.F.			F	l.s.F.
1. (a)	11.84	7.71		(g)	8.22	7.71
(b)	14.04	7.71		(h)	5.26	7.71
(c)	9.82	7.71		(i)	2.84	7.71
(d)	1.26	7.71		(j)	5.05	7.71
(e)	4.27	6.61		(k)	11.37	7.71
(f)	10.29	5.99		(l)	15.47	7.71

2. (a)	2.65			(e)	1.77
(b)	2.51			(f)	3.46
(c)	16.00			(g)	1.95
(d)	21.33			(h)	1.15

3. (a)	0.15			(h)	6.65
(b)	15.09			(i)	15.28
(c)	4.80			(j)	13.10
(d)	12.60			(k)	4.28
(e)	4.80			(l)	18.08
(f)	0.75			(m)	6.05
(g)	1.55				

4. $F = 5.02$; l.h.s.$F = 2.51$.

	F				F
5. (a)	5.40		**6.** (a)	27.00
(b)	0.16		(b)	1.11
(c)	6.96		(c)	14.55
(d)	3.20		(d)	5.00
(e)	1.95		(e)	34.28
(f)	1.95		(f)	48.00
(g)	9.62		(g)	46.88
(h)	10.09		(h)	44.26
(i)	31.19		(i)	22.91
(j)	9.24		(j)	7.44
(k)	2.73		(k)	10.42
(l)	75.29		(l)	474.92
(m)	3.86		(m)	9.85
(n)	1.54		(n)	9.69

B. Problems

7. In reporting electric power production (millions of kilowatt-hours) for a certain state, the Federal Power Commission made a preliminary report (1), a corrected report (2), and a final corrected report (3). Did these reports change significantly, as indicated by the following data for 1937?

Month	(1)	(2)	(3)
Jan.	148	146	148
Feb.	146	145	145
Mar.	171	171	172
Apr.	172	174	174
May	174	174	176
June	178	178	178
July	165	164	164
Aug.	153	153	153
Sept.	135	134	134
Oct.	142	142	142
Nov.	137	137	138
Dec.	138	138	139

8. Time and motion studies were carried on for several months by a manufacturing concern in Seattle, Washington. In one department, where three laborers were employed, a record (1) was made previous to the studies, of the average hourly output for each worker for a period of one week. A similar record (2) was made after the research had been completed and the results had been applied. Was there any significant improvement?

Workers	(1)	(2)
A	2	8
B	2	6
C	5	7

9. A certain corporation had six separate factories scattered through Iowa, Illinois, and Minnesota. Efficiency tests were given to workers in each of the factories in an attempt to find whether there was any significant difference

from one factory to the next. On the basis of the data collected, as shown below, determine whether there was any significant variability. (Data simplified.)

Factory					
A	B	C	D	E	F
1	1				
2	3	7	6		
4	3	9	9	12	12
5	6	10	13	14	18
5	7	11	15	14	20
7	10	13	17	20	26

10. The X Motor Sales Company handles the retail agency for three leading low-priced automobiles, I, II, and III. Four salesmen, A, B, C, and D, handle the entire selling work for the concern. The following data represent the number of cars sold in each of three months during 1939. Compute the significance of the variability in the popular demands for the cars in each of the three months. Would you say that the sales area obviously preferred any one automobile to the rest?

Salesmen	March			July			December		
	I	II	III	I	II	III	I	II	III
A	4	2	3	3	2	4	1	4	4
B	2	5	5	4	5	6	3	4	5
C	1	3	5	3	6	6	2	3	4
D	1	2	3	6	7	8	2	1	3

CHAPTER XX

ELEMENTARY PROBABILITY

Reference has been made in numerous connections throughout earlier chapters to the chance probabilities inherent in data that may be subjected to statistical analysis. Thus, in the early discussion of sampling, it was noted that the "probable" or "standard" errors of various measures derived from samples might be estimated, primarily as a means of appraising the reliability of the measures so derived. Again, in the preliminary consideration of correlation, attention was directed to the possibility of evaluating the coefficient thus obtained by reference to the chance probabilities of such correlation, and similar means of appraising the more complicated measures of correlation and of other variance analysis have been considered in immediately preceding chapters. Throughout all these applications, reference to estimates of chance probabilities implies an understanding of the characteristics of various "chance" distributions, of distributions, in other words, whose magnitudes and frequencies reflect the probabilities of chance occurrences as defined by the nature of the data.

Most frequently, though not by any means universally, distributions representing these chance probabilities take the general form represented by a symmetrical, bell-shaped curve which is described as the "normal curve" or the curve of a "normal distribution." In other cases, notably in the distribution of F, the basic distribution with which comparisons are made is skewed, because the probabilities inherent in the data are not so simple as those reflected in the normal distribution.

In this chapter, attention is directed to some of the elementary characteristics of such chance distributions, including both the usual normal distribution and some others, and to

489

methods of estimating probabilities from the known characteristics of the variables involved.

Binomial probabilities and distributions.—As has been said, the most frequently used statement of probabilities is that which is represented by the Bernoulli or binomial distribution, which in its limiting form is the curve of normal probability. This distribution describes in its series of frequencies for successive magnitudes, the terms of an expanded binomial such as $(1 + a)^n$. Thus, for three classes in the distribution, the frequencies would appear in the proportions indicated by the coefficients of a in the simple expansion

$$(1 + a)^2 = 1a^0 + 2a^1 + 1a^2$$

i.e., as 1, 2, 1. The proportionate frequencies for larger numbers of classes may be as readily derived from other expansions of the same binomial, as, for instance,

For 3 classes, $(1 + a)^2 = 1 + 2a + a^2$
For 4 classes, $(1 + a)^3 = 1 + 3a + 3a^2 + a^3$
For 5 classes, $(1 + a)^4 = 1 + 4a + 6a^2 + 4a^3 + a^4$
For 6 classes, $(1 + a)^5 = 1 + 5a + 10a^2 + 10a^3 + 5a^4 + a^5$
    etc.

In these expansions the successive terms, from 1 (that is, a^0) to the highest powers of a, represent the classes, while the coefficients preceding each a represent the frequencies. Thus the frequencies in the last expansion are, successively, 1; 5; 10; 10; 5; 1. If the binomial is written with two letters, a and b, and raised to the fourth power as

$$(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$

the successive terms contain the descending and ascending powers of these letters, but the frequencies are the same as before. Typical frequencies for successive powers of the binomials, together with other constant characteristics of such distributions, are listed in Table 20·1.

Distributions thus arranged represent in theoretical form the typical ratios of frequencies in many ordinary types of distributions. For example, successive sizes of suits or hats or shoes might be expected to follow some such distribution. Similarly,

if cities were classified according to the number of filling stations or hot-dog stands or beauty parlors in each of them, some such distribution might appear, and a similar result might follow classification of members of a class in statistics on the basis of their grades in the course or their all-university grade averages. Sometimes, notably for instance where classifications are based

TABLE 20·1

FREQUENCIES AND RELATED CHARACTERISTICS OF THE BINOMIAL DISTRIBUTION

(Classes with unit intervals assumed)

Binomial and power (n)	Classes (n + 1)	Frequencies	Σf, or $N = 2^n$	Variance $\sigma^2 = n/4$
$(a + b)^1$	2	1 : 1	2	1/4 = 0.25
$(a + b)^2$	3	1 : 2 : 1	4	2/4 = 0.50
$(a + b)^3$	4	1 : 3 : 3 : 1	8	3/4 = 0.75
$(a + b)^4$	5	1 : 4 : 6 : 4 : 1	16	4/4 = 1.00
$(a + b)^5$	6	1 : 5 : 10 : 10 : 5 : 1	32	5/4 = 1.25
$(a + b)^6$	7	1 : 6 : 15 : 20 : 15 : 6 : 1	64	6/4 = 1.50
$(a + b)^7$	8	1 : 7 : 21 : 35 : 35 : 21 : 7 : 1	128	7/4 = 1.75
$(a + b)^8$	9	1 : 8 : 28 : 56 : 70 : 56 : 28 : 8 : 1	256	8/4 = 2.00
$(a + b)^9$	10	1 : 9 : 36 : 84 : 126 : 126 : 84 : 36 : 9 : 1	512	9/4 = 2.25
$(a + b)^{10}$	11	1 : 10 : 45 : 120 : 210 : 252 : 210 : 120 : 45 : 10 : 1	1024	10/4 = 2.50
$(a + b)^{11}$	12	1 : 11 : 55 : 165 : 330 : 462 : 462 : 330 : 165 : 55 : 11 : 1	2048	11/4 = 2.75

on incomes, distributions are found to be skewed and thus reflecting probabilities different from those represented by the normal curve. Such distributions, both skewed and symmetrical, may be described in mathematical terms as expressions of chance probability. A simple example of such variability is represented by the results obtained in tossing coins, and that illustration may serve to introduce the general principles inherent in such chance variation.

Simple coin-tossing.—Suppose, for instance, that a single coin is "flipped." It is apparent that the probability of its landing head up is 1 out of 2 or $\frac{1}{2}$. Then suppose that 2 coins are tossed together in several successive throws, and a record is kept of the number of heads turned up at each throw. It is obviously possible that any throw may have any one of the following three results: no heads; 1 head; 2 heads. The betting odds for each of these possibilities may readily be calculated, but in order to do so, each of the coins should be

numbered so that they can be distinguished when they are thrown together.[1] Then the possible *permutations*, that is, *specific arrangements* of the two coins in any throw, are:

PERMUTATION	COIN 1	COIN 2
1	Head	Head
2	Head	Tail
3	Tail	Head
4	Tail	Tail

If the specific designations of the coins are ignored, it will be noted that the four *permutations* reduce to three *combinations*, for the second and third become one combination. It may be concluded, therefore, that there is 1 chance in 4 of throwing 0 heads (2 tails), 2 chances in 4 of throwing 1 head (1 tail), and 1 chance in 4 of throwing 2 heads (0 tails). The probabilities for throwing tails may be similarly described.

As is suggested by this illustration, a general rule with respect to the number of probable permutations to be expected from successive chances or throws, where each throw involves one or more coins, may be stated as follows:

1. The number of possible permutations (frequencies) for each throw is equal to the product of the chance possibilities of each coin (2), or 2 raised to the power indicated by the number of coins. It follows, as a corollary that:

1a. The probability that any specific permutation will appear is the reciprocal of the number of possible permutations as just calculated.

Thus, for a single coin, the chance possibilities are 2 (head or tail), and the probability of 1 head is, for each toss, $\frac{1}{2}$ while for 2 coins the permutations are $2 \times 2 = 2^2 = 4$, and the probability of any specific permutation (say a head for coin 1 and a tail for coin 2, or the exact reverse) is $\frac{1}{4}$.[2] Similarly, if 3 coins

---

[1] If the order of the coins is reversed, four other permutations will result, but they will not change the probabilities and therefore may be ignored.

[2] It must be recognized that these principles hold only for *specific* permutations in which the 2 coins are carefully distinguished. If the specific designations of the coins are disregarded, reference is made to *combinations* rather than *permutations*, and the probabilities are obviously changed. When n coins are tossed, there are 2^n permutations, but only $n + 1$ combinations. In general, the permutations of any two independent events are the product of the respective possibilities, e.g., a toss of a coin and a die yield $2 \times 6 = 12$ permutations, and the chance of throwing both a head and a six are 1 in 12.

are tossed, there are possible $2 \times 2 \times 2 = 2^3 = 8$ permutations, or a chance of 1 in 8 for each. Or, if 2 dice are thrown, there are possible $6 \times 6 = 6^2 = 36$ permutations, each of which has 1 chance in 36 of appearing. This principle expressing the number of permutations as the product of the successive possibilities is of basic importance in calculating the frequencies of distributions based on chance.

A second principle of chance involves the addition of probabilities. This principle may be stated as follows:

2. If under set conditions the probability of one result (permutation or combination) occurring is known, and the probability of a second result occurring is also known, then the probability that either one or the other will occur is the sum of these two probabilities.

This principle may be readily illustrated in tossing 2 coins, by estimating the chances of throwing 2 like coins, i.e., 2 heads or 2 tails. The probability of throwing 2 heads at a toss is 1 in 4, and the probability of throwing 2 tails is likewise 1 in 4. Hence the probability of throwing either one or the other, that is, 2 coins alike, is $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. In throwing 3 coins, the chances of throwing 3 heads at any toss is $\frac{1}{8}$, and of throwing either 3 heads or 3 tails, that.is, 3 like coins, is $\frac{1}{8} + \frac{1}{8} = \frac{1}{4}$. In throwing 2 dice the chances of throwing aces are $(\frac{1}{6})^2$ and the chances of throwing 2 alike, that is, 1 and 1, 2 and 2, etc., are $\frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{6}$.

On the basis of these two principles of multiplying and adding probabilities, it is possible to develop independently the frequencies of the distributions listed in Table 20·1. The elementary case of heads (H) and tails (T) in tossing 2 coins, $(H + T)^2$, has already been considered, and it is obvious that the probabilities are expressed by the following frequency distribution:

HEADS	FREQUENCIES (chances in 4)
0	1
1	2
2	1
	——
	4

The chances involved in throwing 3 coins may be similarly expressed. If the coins are numbered, 8 permutations are possible, as indicated in Example 20·1. But if the specific

EXAMPLE 20·1

PROBABILITIES IN TOSSING 3 COINS AT A TIME

(H = heads; T = tails)

Permutations Possible throws			Combinations Total in each throw	
Coin number: 1	2	3	H	T
1st throw	H H H		3	0
2nd throw	H H T		2	1
3rd throw	H T H		2	1
4th throw	H T T		1	2
5th throw	T H H		2	1
6th throw	T H T		1	2
7th throw	T T H		1	2
8th throw	T T T		0	3

Summary (exponents indicate number of H or T):

$$(H + T)^3 = H^3 + 3H^2T + 3HT^2 + T^3$$

or, in terms of heads only:

$$(1 + H)^3 = 1 + 3H + 3H^2 + H^3$$

where 1 means 1 throw of no heads.

designation of the coins is ignored, the number of combinations reduces to 4, and the following distribution results:

HEADS	FREQUENCIES (chances in 8)
0	1
1	3
2	3
3	1
	8

In the same way the probabilities in throwing 4 or more coins at each throw may be calculated. But it is easier to express the results by following a general rule than by analyzing the permutations in each case. The general rule in its simplest form may be seen by reference to Table 20·1 where any given frequency may be derived as the sum of the two nearest frequencies in the preceding line.[1] This rule, however, is not convenient in application to multiclass expansions, since it requires a build-up from smaller expansions. The more general rule for the independent expansion of a binomial may therefore be stated as the frequency distribution $(0.5 + 0.5)^n$, where n is the number of coins tossed, 2^n is the number of permutations (frequencies), and $n + 1$ is the number of combinations (classes). It may be shown that the variance of such a distribution is $n/4$.

Algebraic rule for binomial expansions.—The general algebraic rule for expanding a binomial representing a sequence of probabilities may be readily stated. In so doing it is convenient to distinguish the separate elements entering into the terms. These elements have been designated a and b, or H and T, but further analysis is facilitated if they are called p and q. As has been noted, the first letter appears in descending powers, beginning with the power of the expansion (n) or the number of classes less 1. The second letter appears in ascending powers, beginning with the zero power of the letter, which equals unity. The frequencies begin with 1, the next is n, and the rest may be obtained by successive fractional multipliers with descending numerators and ascending denominators. For example, if $n = 4$, successive frequencies may be found by multiplying 1 cumulatively by $\frac{4}{1}$, $\frac{3}{2}$, $\frac{2}{3}$, and $\frac{1}{4}$, producing frequencies of 1, 4, 6, 4, and 1. If $n = 5$, the factors applied cumulatively to the first frequency, 1, are $\frac{5}{1}$, $\frac{4}{2}$, $\frac{3}{3}$, $\frac{2}{4}$, and $\frac{1}{5}$, and the frequencies therefore are 1, 5, 10, 10, 5, and 1. The process may be visualized by setting up the elements separately, and then combining them into the required terms, as is done in Example 20·2. For convenience in later calculations the class marks (m or X) are

[1] The data of this table are frequently referred to as *Pascal's triangle*. As stated in the table, the top row represents the coefficients of the expansion to the first power, the second row the square, etc.

taken as the sequence 0, 1, 2 . . . n, though their interpretation may sometimes suggest a reverse order. In terms of coin tossing, p and q without designated values may be interpreted as heads and tails, or tails and heads, replacing H and T of previous illustrations.

<div style="text-align:center">

EXAMPLE 20·2

EXPANSION OF THE BINOMIAL $(p + q)^n$ WHERE $n = 4$

</div>

m	0	1	2	3	4
f and f multiplier	1	$\frac{4}{1}$　4	$\frac{3}{2}$　6	$\frac{2}{3}$　4	$\frac{1}{4}$　1
Powers of p Powers of q Terms	p^4 q^0 p^4	p^3 q^1 $4p^3q$	p^2 q^2 $6p^2q^2$	p^1 q^3 $4pq^3$	p^0 q^4 q^4

Hence

$$(p + q)^4 = p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4$$

Any term (designated by m) may be calculated as

$$\text{term } m = \frac{n!}{(n - m)!\, m!}\, p^{n-m}q^m$$

where ! means the series of factors 1, 2, 3, etc., to the number indicated. For example, term number $m = 3$ is

$$\frac{n!}{(n - m)!\, m!}\, p^{n-m}q^m = \frac{1 \times 2 \times 3 \times 4}{(1)(1 \times 2 \times 3)} = 4pq^3$$

Note that 0! implies omission, not a zero factor, while p^0 and q^0 each equals 1.

In order to utilize binomial expansions as representative of typical distributions of business data, it is necessary to go one step further in interpreting the letters comprising the binomials. Suppose, for example, that p is taken to represent the probability that for any 1 coin tossed a head will turn up, and q that it will not. Then $p = q = \frac{1}{2}$, and $p + q = 1$, or certainty that either head or tail will turn up. If the expansion is carried

out with these values of p and q, following the pattern of Example 20·2, the powers of these letters become:

Powers of p:	1/16	1/8	1/4	1/2	1
Powers of q:	1	1/2	1/4	1/8	1/16

while the frequencies are as before, namely,

Frequencies:	1	4	6	4	1

If these three elements are combined by multiplying them, term by term, the resulting probabilities (P) are

Heads (H):	4	3	2	1	0
Tails (T):	0	1	2	3	4
P:	0.0625	0.2500	0.3750	0.2500	0.0625

and these are the probabilities (625 in 10,000; 2,500 in 10,000, etc.) that each given combination will appear, namely (if p = heads and q = tails), all heads; 3 heads, 1 tail; 2 heads, 2 tails; 1 head 3 tails; and all tails. In this case, therefore, the interpretation is the same as before. However, the sum of these probabilities is now 1, or certainty that some one of these combinations will turn up when 4 coins are tossed together.

Skewed binomial expansions.—The use of p and q to express alternative probabilities, as just explained, makes it possible to express probabilities in cases where the chances are something other than 1 in 2. Suppose, for example, that 3 dice are thrown together, and the probabilities of throwing aces (ones) is in question. For 1 die the chance of turning up an ace is $\frac{1}{6}$, and the chance of turning up something else is $\frac{5}{6}$. Then for a set of 3 dice the probabilities are expressed by the binomial:

Probabilities of aces, 3 dice: $(p + q)^n = (\frac{1}{6} + \frac{5}{6})^3$.

If this binomial is expanded according to the rules of Example 20·2, the probabilities for 1, 2, or 3 aces may be found as:

No. of Aces:	3	2	1	0
p:	$(1/6)^3$	$(1/6)^2$	$(1/6)$	1
q:	1	$(5/6)$	$(5/6)^2$	$(5/6)^3$
f:	1	3	3	1

Product of p, q, and f elements for each column:

P: $1/216 + 15/216 + 75/216 + 125/216 = 216/216$

or: $0.00463 + 0.06945 + 0.34722 + 0.57870 = 1.00000$

Thus there is a probability of only 1 in 216, or 463 in 100,000, that all aces will be thrown at any 1 toss of 3 dice. The chances for 2 aces or 1 ace are increasingly better, while the chances of throwing no aces at a toss are a little better than $\frac{1}{2}$. By successive additions it may be seen that there are $\frac{1}{216} + \frac{15}{216} = \frac{16}{216}$ chances of throwing at least 2 aces, and $\frac{1}{216} + \frac{15}{216} + \frac{75}{216} = \frac{91}{216}$ chances of throwing at least 1 ace.

Illustrations from business.—Such probabilities may sometimes be applied to business situations. For example, in a certain stable business environment, there had been strong competition in the field of ordinary restaurants, and the annual rate of failure of new ventures was 1 in 6. The chance of succeeding was therefore 5 in 6. If at a given time 4 new restaurants are started, the number not being unusual, the chances of success or failure during the ensuing year may be expressed by the expansion of

$$(p + q)^n = (\tfrac{1}{6} + \tfrac{5}{6})^4$$

This binomial, expanded according to the form just described, gives these results: [1]

Successes (m):	0	1	2	3	4
Failures:	4	3	2	1	0
P:	1/1296	20/1296	150/1296	500/1296	625/1296
or:	0.001--	0.015	0.116	0.386	0.482

These probabilities mean that, of the 4 restaurants in question,

----

[1] When n is large, say 8 or more, binomials are best expanded by the use of logs of p and q, with p and q first written as decimals and multiplied by some convenient power of 10 to avoid negative logs. In the final probabilities, the multiplier may easily be eliminated by a suitable placing of the decimal. The logs of the expression $p^{n-m}q^m$ with p and q modified as indicated may readily be found. The first log term will be n times log p. Others may be formed by successive subtractions of the quantity, (log p − log q). The antilogs may then be taken, multiplied by the frequencies, and suitably pointed off. Or the frequencies may also be expressed as logs (log f_1 = log 1; log f_2 = log f_1 + log n − log 1; log f_3 = log f_2 + log $\overline{n-1}$ − log 2, etc.) and added to the pq logs.

the chances for the year are only 1 in 1,000 that none will succeed (4 failures), 15 in 1,000 that only 1 will succeed (3 failures), 116 in 1,000 that 2 will succeed (2 failures), 368 in 1,000 that 3 will succeed (1 failure), and 482 in 1,000 that all will succeed (no failures). Cumulated in reverse order, the probabilities also indicate that there are 868 chances in 1,000 that at least 3 will live through the year, 984 in 1,000 that at least 2 will survive, and 999 in 1,000 that at least 1 will survive.

In many problems involving the use of binomial distributions it is desirable to know the mean and variance. These may, of course, be found by the usual methods of computation as applied to a frequency distribution, but they are much more conveniently found by the formulas [2] $M = nq$ and $\sigma^2 = npq$. These formulas refer to the m scale 0, 1, 2, ... n, and the frequencies include the appropriate p and q values. For example, in the distribution quoted above, namely,

m:	0	1	2	3	4
f:	1	20	150	500	625

where $(p + q)^n = (\frac{1}{6} + \frac{5}{6})^4$, the mean is

$$M = nq = 4 \times \tfrac{5}{6} = 3.3\underline{3}$$

$$\sigma^2 = npq = 4 \times \tfrac{1}{6} \times \tfrac{5}{6} = 0.5\underline{5}$$

These results will be found to agree with those calculated by the usual procedure.

A more important use of the binomial distribution may be found in determining the relative numbers of sizes of shoes, hats, garments, or other articles to be purchased in laying in a stock. For many items, of course, the classification of sizes has already been reduced to a routine as a result of experience, and no special calculation is necessary. But it might happen that comparatively little is known regarding the size distribution of a new line of merchandise. As an illustration may be cited a new line which came in sizes 2, 4, 6, 8, and 10. Little was known of the relative number of each size sold, but for a

[2] $Mo = M + q - 0.5$, approximately.

considerable sample of sales, a mean size of 5.2 had been computed. Assuming a chance distribution, could the relative number of sizes suitable for a stock be estimated?

The frequencies of the skewed binomial distribution, $(p + q)^n$, where the mean is nq, may be calculated by first determining q as

$$q = (M - m_0) \div in = (5.2 - 2) \div (2 \times 4) = 0.4$$ where m_0 is the smallest size, 2, and M the mean, 5.2. Then $p = 1 - q = 1 - 0.4 = 0.6$, and the required frequencies are

$$
\begin{aligned}
(p + q)^n &= (0.6 + 0.4)^4 \\
&= 0.1296 + 0.3456 + 0.3456 + 0.1536 + 0.0256
\end{aligned}
$$

or, with rounded frequencies, the distribution is

Size (m):	2	4	6	8	10
Frequencies:	13	35	34	15	3

Many problems of this type arise in business data. Often the frequencies are calculated according to the so-called normal curve, which will be given consideration in this connection.

FITTING THE NORMAL CURVE

Theoretical frequencies may be fitted to data, not only by means of binomial distributions as just described, but also by means of the normal curve (cf. pages 582–585). When this is done, allowance for skewness is a difficult matter, not commonly attempted, and its consideration has here been relegated to the Appendix (cf. pages 516–519). But, assuming that the data do not involve an important degree of skewness, an approximate fitting is described in the following paragraphs, where it is applied to a problem similar to one considered in connection with binomial distributions.

To illustrate use of the normal curve as a basis for estimating probable frequencies, it may be assumed that a retail dealer who is introducing a certain brand of shoes finds that the first 75 pairs sold are distributed, according to size, as follows:

Size	Number of Pairs Sold
m	f
$6\frac{1}{2}$	0
7	3
$7\frac{1}{2}$	9
8	17
$8\frac{1}{2}$	22
9	15
$9\frac{1}{2}$	7
10	2
$10\frac{1}{2}$	0

A chart of this sample distribution indicates that it is fairly normal. Its mean is 8.44, and its standard deviation is 0.6829. The distribution of sizes of 1,000 pairs to be ordered would be based on the proportions of each size included in a normal distribution having the same measures (mean and standard deviation) as the sample.

Normal ordinates.—From the table of ordinates of the normal curve (pages 582–585), the measures of ordinates at each class measure may be secured, and the sum of these z values may be considered as a hypothetical N, from which the estimated frequency at each ordinate may be calculated as its indicated proportion of the total. Since the unit of measurement on the *abscissa* of the normal curve is the standard deviation, it is first necessary to express each mid-point as a deviation (x) from the mean of the distribution ($x = X - M_X$), and then to describe this deviation as a ratio of the standard deviation. Ordinates are designated, in the table, by this ratio, $x \div \sigma_X$. Thus the appropriate designation for the ordinate at size 6.5 would be

$$\frac{x}{\sigma} = \frac{(X - M_X)}{\sigma} = \frac{6.5 - 8.44}{0.6829} = -2.84$$

The sign may be disregarded and the ordinate read directly from the table as 0.00707. Before this figure can be made significant, it is necessary to measure each of the other ordinates in the same manner. Their sum, Σz, is then noted, after which the ratio of each individual measure to this total is calculated. These ratios may then be applied directly to the N of the dis-

tribution to discover the frequencies at each ordinate. The process is illustrated, step by step, in the columns of Example 20·3.[1]

EXAMPLE 20.3

ESTIMATING THEORETICAL FREQUENCIES ON THE BASIS OF ORDINATES [1]

Data: Assumed normal distribution of shoe sizes, $M = 8.44$, $\sigma = 0.6829$. Theoretical frequencies for 1,000 pairs.

Size	Deviations from mean $X - M_x$	Designation of ordinate	Measure of ordinate	Ratio to total measures	Theoretical frequencies
X	x	$x \div \sigma$	z	$z \div \Sigma z$	
6.5	−1.94	−2.84	0.00707	0.00518	5
7.0	−1.44	−2.11	0.04307	0.03155	31
7.5	−0.94	−1.38	0.15395	0.11275	113
8.0	−0.44	−0.64	0.32506	0.23808	238
8.5	0.06	0.09	0.39733	0.29101	291
9.0	0.56	0.82	0.28504	0.20877	209
9.5	1.06	1.55	0.12001	0.08790	88
10.0	1.56	2.28	0.02965	0.02172	22
10.5	2.06	3.02	0.00417	0.00305	3
			$\Sigma z = 1.36535$		1,000

[1] If the frequencies are determined on the basis of a binomial distribution, as previously explained, they differ from those obtained in the problem above. They are as follows: 5, 37, 123, 231, 273, 205, 97, 26, 3. Since the binomial distribution makes allowance for the small degree of skewness present in the data, it may be regarded as a somewhat preferable method.

It should be indicated that, if the distribution is regarded as continuous, frequencies at ordinates other than those indicated may be calculated in the same manner. But the process of calculation ignores the size of the class interval and is not properly

[1] Sometimes the tables of ordinates express each ordinate as a decimal fraction of the maximum ordinate, the one at the mean. The tabular values of z are thus increased by the factor 2.50663, but they may, nevertheless, be employed in the manner described.

applicable in many instances for this reason. Moreover, as was previously noted, its applicability in this simple form is limited to distributions that closely approximate the normal. A more accurate method, therefore, involves the use of areas of the normal curve between the lower and upper limits of each class, which allows for grouping though not for skewness.

Analysis of area.—Another approach to the same problem makes reference to the area under various portions of the normal curve (see Example 20·4). The portions of this area within various ranges of the mean are known (one standard deviation on each side of the mean, for instance, includes 68.27 per cent of the area, two standard deviations 95.45 per cent, etc.). Hence, any given deviation from the mean may be regarded as a boundary and the area thus defined may be read directly from a table which describes the total area under the curve as unity (column A of the table on page 582). The table, of course, refers to but one-half of the symmetrical curve. Thus, the area included in the range from the mean to one standard deviation (designated 1.00 in the table) is 0.3413. For the whole distribution—one standard deviation on *each side* of the mean—the area would be twice this fraction or 0.6827, a figure already familiar.

In the same way, the area between any two ordinates may be readily determined by subtracting the fraction representing the smaller area from that representing the larger. If, for instance, the ordinates are 1.00 and 2.00, the area included would be $0.47725 - 0.34134 = 0.13591$, which means that approximately 14 per cent of the total distribution appears within these boundaries. This calculation refers to but one side of the symmetrical curve.

The same approach may be applied to each of the classes of a given distribution such as that represented by Example 20·3. Class limits are first located on the abscissa as x/σ, by calculating their ratios (as deviations from the mean) to the standard deviation of the sample. Then, the area between each of these ordinates or boundaries and the mean is read from the table. The net area between the limits of each class is then noted, and this fraction of the total distribution represents the theoretical frequency of the class.

EXAMPLE 20·4

ESTIMATING THEORETICAL FREQUENCIES
ON THE BASIS OF AREA

Data: Assumed normal distribution of shoe sizes (see preceding example).
Theoretical distribution for 1,000 pairs.

Class mark (size)	Class limits	Limits expressed as deviations from the mean	Limits expressed in σ units	Area between given ordinate and maximum ordinate	Net area, each class [1]	Theoretical frequencies
m	L	x	$x \div \sigma$	A		
$6\frac{1}{2}$	6.25	−2.19	−3.21	0.49934		
	6.75	−1.69	−2.47	0.49324	0.00610	6
7					0.03417	34
	7.25	−1.19	−1.74	0.45907		
$7\frac{1}{2}$					0.11532	115
	7.75	−0.69	−1.01	0.34375		
8					0.23349	234
	8.25	−0.19	−0.28	0.11026		
$8\frac{1}{2}$					0.28390	284
	8.75	0.31	0.45	0.17364		
9					0.20934	209
	9.25	0.81	1.19	0.38298		
$9\frac{1}{2}$					0.08959	90
	9.75	1.31	1.92	0.47257		
10					0.02341	24
	10.25	1.81	2.65	0.49598		
$10\frac{1}{2}$					0.00366	4
	10.75	2.31	3.38	0.49964		———
						1,000

[1] In calculating the net area of each class, it is necessary to add the areas represented by the limits of the class in which the mean appears, since this class includes items on both sides of the mean.

The class limits for size $6\frac{1}{2}$ are, for instance, 6.25 and 6.75, which represent deviations of -2.19 and -1.69, respectively. Expressed in σ units (x/σ), the limits become

$$L_1 = \frac{-2.19}{0.6829} = -3.2069; \quad L_2 = \frac{-1.69}{0.6829} = -2.4747$$

respectively. By reference to the table, it will be seen that the area between the first and the mean is 0.49934, while that between the second and the mean is 0.49324. The net area within these limits is, therefore, $0.49934 - 0.49324 = 0.00610$. Applied to the prospective order for 1,000 pairs of shoes, this calculation indicates a theoretical frequency of 6 pairs of this size.

The theoretical frequencies for each of the other sizes involved may be calculated in the same manner, as is indicated in Example 20·4.

THE CHI-SQUARE TEST

Testing for trueness to type.—A final problem requiring both a consideration of probability and a measure of reliability involves comparison of frequency distributions with theoretical distributions in order to determine whether the latter may be assumed to express the nature of the former. To illustrate this type of problem, the following frequency distribution may be cited:

Wages per week	(m) =	\$15	20	25	30	35	40
Number of workers	(f) =	10	14	31	31	13	1

The question is: how closely does this distribution approximate a normal or chance distribution? An analogous binomial normal distribution expressing a common law of chance and involving 6 classes to permit ready comparison with the given frequencies may be calculated by expanding the binomial $(\frac{1}{2} + \frac{1}{2})^5$ to express the coefficients of the terms. As has already been explained, the frequencies are the complex ratio, 1 : 5 : 10 : 10 : 5 : 1 (total 32), which in this case must be increased so that the total is 100, as in the given distribution; that is, they

must be multiplied by $\frac{100}{32}$ = 3.125. The theoretical or expected frequencies (f_e), therefore, become

$$f_e = (1, \quad 5, \quad 10, \quad 10, \quad 5, \quad 1) \quad (3.125)$$

$$= 3.1, \quad 15.6, \quad 31.3, \quad 31.3, \quad 15.6, \quad 3.1$$

Chi-square.—The measurement of the divergences between the actual and the expected frequencies involves the use of a statistic called chi square (χ^2) which is found by the formula

$$\chi^2 = \Sigma \left[f_e \left(\frac{d}{f_e} \right)^2 \right] = \Sigma \frac{(f - f_e)^2}{f_e}$$

where $d = f - f_e$. Since the residuals (d) are expressed as percentages of f_e, and then squared, very small frequencies (below 5, for instance) should be combined to form larger groups. In the case at hand the first two and the last two frequencies are combined, making in all four classes.

The process is illustrated in detail in Example 20·5. The summation of $(f - f_e)^2 \div f_e$, 2.689, is χ^2, which is taken as a measure of the disparity between the frequencies of the data and the frequencies of the corresponding theoretical distribution. The greater the disparity, obviously, the greater χ^2 will become.[1]

By reference to an appropriate table or chart the probability of the data occurring as a chance variation of the theoretical distribution may be determined. Such a measure is presented in graphic form on page 561. It appears that the probability (P)

---

[1] In the calculation of chi square, Yates's correction for continuity is desirable, particularly with one degree of freedom and small values of $f - f_e$. This correction is applied by reducing the numerical value of $f - f_e$ by 0.5, to a limit of zero. To illustrate, in Example 20·5,

f_e	=	18.7	31.3	31.3	18.7
$f - f_e$	=	5.3	−0.3	−0.3	−4.7
Less 0.5	=	4.8	0.0	0.0	−4.2
Squares	=	23.04	0.00	0.00	17.64
Divided by f_e	=	1.23	0.00	0.00	0.94; $\chi^2 = 2.17$

This correction is an approximate compensation for the error that arises when a continuous curve is taken as a measure of a frequency distribution. The error is analogous to that which would occur if the normal curve were taken to represent $(\frac{1}{2} + \frac{1}{2})^n$. If n were small, the error might be very significant (cf. C. H. Goulden, *Methods of Statistical Analysis*, John Wiley & Sons, 1939, pp. 101–104).

EXAMPLE 20·5

THE CHI-SQUARE TEST

Data: Assumed wages, dollars per week in a small factory.

Wage groups (mid-point) m	Number of workers f
$15	10
20	14
25	31
30	31
35	13
40	1
	100

Frequencies in a normal 6-class binomial distribution:

$$\left(\frac{1}{2} + \frac{1}{2}\right)^5 = \frac{(1+1)^5}{2^5} = \frac{(1 + 5 + 10 + 10 + 5 + 1)}{32}$$

Theoretically expected frequencies (f_e) for 100 items:

$$f_e = \frac{100(1 + 5 + 10 + 10 + 5 + 1)}{32}$$

$$f_e = (3.1 + 15.6 + 31.3 + 31.3 + 15.6 + 3.1)$$

f_e (adjusted to avoid small frequencies by combining end classes)

$$= (18.7 + 31.3 + 31.3 + 18.7)$$

Calculation of χ^2:

f	f_e	$(f - f_e)$	$(f - f_e)^2$	$\dfrac{(f - f_e)^2}{f_e}$
24	18.7	5.3	28.09	1.502
31	31.3	−0.3	0.09	0.003
31	31.3	−0.3	0.09	0.003
14	18.7	−4.7	22.09	1.181
Total = χ^2				2.689

Evaluation of χ^2, where $N - m = 4 - 1$:

Figure A5, on page 561, indicates that a χ^2 value of 2.689 for three degrees of freedom has a probability of 0.45 that it might be exceeded by chance.

is approximately 0.45 when χ^2 is 2.689 and the degrees of freedom, that is, $N - m$, is 3. In this case m is taken as 1, since only one criterion—Σf—is used in adjusting the theoretical to the actual frequencies.[1] This means that in 45 samples out of 100 the actual frequencies might by chance diverge further from the theoretical distribution. Hence, it may be concluded that the actual distribution is not significantly different from the binomial distribution assumed to represent its type.

If, however, χ^2 had indicated a probability of only 0.05, the divergence of the actual from the theoretical would doubtless be significant, and some other type of distribution should be assumed. This conclusion is practically certain with a probability of 0.01. But a stricter test is indicated.

Chi square and phi.—As a further example of chi square, reference may be made to the fourfold table previously cited as a type of correlation procedure (page 362), the data of which are as follows:

	FAILED	SUCCEEDED	TOTALS
Trained	$a = 52$	$b = 25$	$g = 77$
Untrained	$c = 95$	$d = 23$	$h = 118$
Totals	$e = 147$	$f = 48$	$N = 195$

It was stated that, in this case, the number of degrees of freedom for $N\phi^2$, that is, chi square, is $4 - 3 = 1$. The hypothesis sets up as the origin of the deviations, (f_e), chance values of a, b, c, and d having the same grand total as the data, and the same column and row totals. Computed $a = (147 \times 77) \div 195 = 58.05$, and b, c, and d are similarly found to be 18.95, 88.95, and 29.05, respectively, their total being 195. These items are chosen because they represent a null hypothesis, or no apparent effect of training, and chi square is used as a means of determining whether the actual results differ significantly from this hypothetical standard, whether, in other words, training produces frequencies differing significantly from those that would be expected if it were completely ineffective. The computation

---

[1] If a skewed binomial had been employed, adjusted to the data by both Σf and M, then m would have been 2. If a normal distribution had been fitted, adjusted to Σf, M, and σ, then m would have been 3.

may be made on the basis of $f - f_e$ as in Example 20·5, or by algebraic short cuts as previously explained. The measurement of mean square contingency (page 427) is merely an expansion of the same principle.[1]

The *general* conditions to which this type of analysis must conform are, first, that the grand total is fixed (in this case, 195), and second, that the totals by columns and rows agree with the corresponding data totals. The theoretical frequencies, expected by the null hypothesis, are obviously adapted to the data by three criteria: the grand total, one row total, and one column total, from which the other totals may be deduced. There is, therefore, only one degree of freedom. This conclusion may be confirmed by the fact that if only one item is known, within the given frame of reference, the others are fixed by the frame of totals.

Conclusion.—There are, of course, many applications of the theory of probability in the determination of the significance of statistical measures. In preceding paragraphs, the elementary principles of this theory have been described. In conclusion, it should be emphasized that the theory of sampling assumes the collection of a comparatively small sample taken from virtually unlimited populations or universes of data. It is obvious that, to the extent that a sample approaches in size the whole from which it is drawn, its reliability is increased. If, for

---

[1] Directly computed:

$$\chi^2 = \Sigma \frac{(f - f_e)^2}{f_e} = \frac{(52 - 58.05)^2}{58.05} + \frac{(25 - 18.95)^2}{18.95} + \frac{(95 - 88.95)^2}{88.95} + \frac{(23 - 29.05)^2}{29.05}$$

$$= 0.63053 + 1.93153 + 0.41150 + 1.25998 = 4.234$$

Or, computed by ϕ^2:

$$\chi^2 = N\phi^2 = N\left(\frac{a^2}{eg} + \frac{b^2}{fg} + \frac{c^2}{eh} + \frac{d^2}{fh} - 1\right)$$

$$= 195\left(\frac{52^2}{147 \times 77} + \frac{25^2}{48 \times 77} + \frac{95^2}{147 \times 118} + \frac{23^2}{48 \times 118} - 1\right)$$

$$= 195(0.23889 + 0.16910 + 0.52029 + 0.09340 - 1) = 4.228$$

5 per cent probability, $\chi^2 = 3.841$; 1 per cent, $\chi^2 = 6.635$.

If Yates's correction is applied, $\chi^2 = 3.56$.

example, a sample of 900 cases were taken from a population that totaled only 1,000, the probability that the mean of the sample accurately represented the total would be high, since $\sigma_m^2 = (\sigma^2/N)(1 - N/N_u)$, where N_u is the total population.

Complex mathematical procedures may be derived for taking into account this feature of sampling, but they are not often used, mainly because of the fact that the term "population" is, in a sense, elastic. For example, in measuring a certain characteristic of the population of a given city, the statistician might obtain a large sample, thus providing a very reliable measure for that city. But the same sample might be considered fairly representative of similar cities in the state or section of the country, so that it becomes but a small (and scarcely random) sample of the whole population from which it is drawn. The relationship of the size of the sample to that of the population with respect to which conclusions are to be drawn must, therefore, be clearly recognized and taken into account.

The theory of probability also assumes that variations are normally distributed in the population from which it is drawn. Actually, many of the "populations" encountered in all social sciences, and to a considerable extent in the biological and physical sciences as well, are not so distributed, but are more likely to represent "logarithmic normals," that is, they take the normal, bell-shaped form only when the logarithms of the X values are plotted or when the same values plus some undetermined constant are utilized in logarithmic form, or they exhibit still different characteristics. Frequently it is possible to transfer a distribution to a logarithmic scale, so that it assumes an approximately normal form, after which the various measures of probability may be readily applied. Although this is not the regular procedure, the principle involved may well be kept in mind in order to avoid misinterpretation of probability calculations.

For example, in an extremely skewed distribution of the logarithmic type, the statement that practically the entire distribution will be found within the range $M \pm 3\sigma$ may be quite erroneous. In reality, both limits thus estimated may be well below actual limits, although the estimate of the standard error

of the mean will probably not be so badly distorted. Because of the frequency with which such distributions appear, it is likely that in time the procedures used in the measurement of reliability will be revised and extended to take account of the type form of distributions to which they may be applied.

READINGS

(Also see special and general references, pages 591–597.)

BATEN, W. D., *Elementary Mathematical Statistics*, New York, John Wiley & Sons, 1938, Chapter IV.

BERKSON, JOSEPH, "Some Difficulties of Interpretation in the Chi-Square Test," *Journal of the American Statistical Association*, 33 (203), September, 1938, pp. 526–536.

CAMP, BURTON H., "Further Interpretations of the Chi-Square Test," *Journal of the American Statistical Association*, 33 (203), September, 1938, pp. 537–542.

DEMING, W. EDWARDS, "The Chi-Test and Curve Fitting," *Journal of the American Statistical Association*, 29 (188), December, 1934, pp. 372–383.

FRY, THORNTON C., "The χ^2-Test of Significance," *Journal of the American Statistical Association*, 33 (203), September, 1938, pp. 513–525.

GOULDEN, C. H., *Methods of Statistical Analysis*, New York, John Wiley & Sons, 1939, Chapters III, IX, and X.

IRWIN, J. O., "Note on the χ^2 Test of Goodness of Fit," *Journal of the Royal Statistical Society*, 92 (2), 1929, pp. 264–266.

KENNEY, JOHN F., *Mathematics of Statistics*, New York, D. Van Nostrand Company, 1939, Part I, Chapter VI; Part II, Chapters I, VII, VIII.

MILLS, F. C., *Statistical Methods* (revised), New York, Henry Holt & Co., 1938, Chapter XIII.

MUDGETT, BRUCE D., and WOLFE, F. E., "The Application of the Theory of Sampling to Successive Observations not Independent of Each Other," *Proceedings of the American Statistical Association*, 24 (165A-Supplement), March, 1929, pp. 108–117.

NAGEL, ERNEST, *Principles of the Theory of Probability*, Chicago, University of Chicago Press, 1939, 80 pages.

NEYMAN, J., and PEARSON, E. S., "Further Notes on the χ^2 Distribution," *Biometrica*, 22 (3–4), May, 1931, pp. 298–305.

TRELOAR, ALAN E. *Elements of Statistical Reasoning*, New York, John Wiley & Sons, 1939, Chapters VI and XII.

WEARER, WARREN, "The Reign of Probability," *Scientific Monthly*, 31 (5), November, 1930, pp. 457–466.

WILSON, EDWIN B.; HILFERTY, MARGARET M.; MAHER, HELEN C., "Goodness of Fit," *Journal of the American Statistical Association*, 26 (176), December, 1931, pp. 443–448.

EXERCISES AND PROBLEMS

A. EXERCISES

1. (a) Express the sampling distribution of heads in throwing a set of 3 coins.

(b) What is the probability of no heads in any one throw? Of 3 heads in any one throw?

(c) What is the probability of throwing 1 or more heads?

2. (a) What is the sampling distribution of total spots in each throw, if 2 dice are thrown?

(b) What is the probability of 10 or more spots in any one throw?

3. Employing the normal curve as an approximation, estimate the probability that 40 heads or more will turn up in a throw of 64 coins. (NOTE: Find the σ's from 32 to 39.5, where $\sigma^2 = npq$.)

4. Expand the following distributions, where $X = 0, 1 \ldots n$, and express the probability of occurrence of 1 or more.

(a) $Y = (0.8 + 0.2)^4$.

(b) $Y = (0.7 + 0.3)^5$.

(c) $Y = (0.6 + 0.4)^6$.

5. Are the following distributions significantly different from the binomial $(0.5 + 0.5)^n$ as measured by the 5 per cent level of chi square?

(a) $f = 20; 30; 20; 10$.

(b) $f = 7; 36; 50; 26; 9$.

(c) $f = 14; 18; 120; 64; 20; 20$.

ANSWERS TO EXERCISES

1. (a) $(0.5 + 0.5)^3$. $H = 0, 1, 2, 3$. $f = 1, 3, 3, 1$. $f\% = 12.5, 37.5, 37.5, 12.5$.

(b) No heads: 12.5%; 3 heads: 12.5%.

(c) $37.5\% + 37.5\% + 12.5\% = 87.5\%$.

2. (a) Permutations $(X + Y)$:

X Y	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Combinations:

Spots: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

f: 1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1. $N = 36$

$f\%$: 2.78, 5.56, 8.33, 11.11, 13.89, 16.66, 13.89, 11.11, 8.33, 5.56, 2.78. $\Sigma\% = 100$

(b) $8.33\% + 5.56\% + 2.78\% = 16.67\%$.

3. $\sigma = 64 \times 0.5 \times 0.5 = 4$; $x = 39.5 - 32 = 7.5$.

$x/\sigma = 7.5 \div 4 = 1.875$; by table, area from $x/\sigma = 0$ to $x/\sigma = 1.875 = 0.47$.
Including negative area, probability below $40 = 0.50 + 0.47 = 0.97$.
Probability at 40 or above $= 1.00 - 0.97 = 0.03 = 3\%$.

4. (a) $Y = \dfrac{1}{10^4}(8 + 2)^4$

Powers of p.	8^4	8^3	8^2	8^1	8^0	
Powers of q.	2^0	2^1	2^2	2^3	2^4	
Frequencies.	1	4	6	4	1	
Expansion..	4,096	4,096	1,536	256	16.	$\Sigma = 10,000$
X..........	0	1	2	3	4	

$P(1 - 4) = 40.96\% + 15.36\% + 2.56\% + 0.16\% = 59.04\%$

(b) $Y = \dfrac{1}{10^5}(7 + 3)$

Powers of p...	7^5	7^4	7^3	7^2	7^1	7^0	
Powers of q...	3^0	3^1	3^2	3^3	3^4	3^5	
Frequencies...	1	5	10	10	5	1	
Expansion....	16,807	36,015	30,870	13,230	2,835	243	$\Sigma = 100,000$
X...........	0	1	2	3	4	5	

$P(1 - 5) = 36.015\% + 30.870\% + 13.230\% + 2.835\% + 0.243\% = 83.193\%$

(c) $Y = \dfrac{1}{10^6}(6 + 4)$

Powers of p.	6^6	6^5	6^4	6^3	6^2	6^1	6^0	
Powers of q.	4^0	4^1	4^2	4^3	4^4	4^5	4^6	
Frequencies.	1	6	15	20	15	6	1	
Expansion..	46,656	186,624	311,040	276,480	138,240	36,864	4,096	
								$\Sigma = 1,000,000$
X.........	0	1	2	3	4	5	6	

$P(1 - 6) = 100.0000\% - 4.6656\% = 95.3344\%$.

5. (a) Significantly different:

f	20	30	20	10
f_e	10	30	30	10
d	10	0	-10	0
d^2	100	0	100	0

d^2/f_e	10.0	0	3.333	0
x^2	13.333		$n = 3$	
5% level	7.815			
1% level	11.341			

(b) Not significantly different:

f	7	36	50	26	9
f_e	8	32	48	32	8
d	−1	4	2	−6	1
d^2	1	16	4	36	1
d^2/f_e	0.125	0.5	0.083	1.125	0.125

$\chi^2 = 1.958$ $n = 4$
5% level $= 9.488$

(c) Significantly different:

f	14	18	120	64	20	20
f_e	8	40	80	80	40	8
d	6	−22	40	−16	−20	12
d^2	36	484	1,600	256	400	144
d^2/f_e	4.5	12.1	20	3.2	10	18

$\chi^2 = 67.8$ $n = 5$
1% level $= 15.1$

B. PROBLEMS

6. In a certain class of stores, in relatively normal years, the chances of surviving are expressed by the ratio 4 : 1. In a given group of 7 stores, what are the relative chances of the number surviving at the end of a year?

7. A random sample of 100 voters out of a large electorate showed that 65 favored a certain measure then under discussion. Is a favorable majority highly probable? (Disprove the hypothesis of a 50–50 vote, with npq of the binomial $(0.5 + 0.5)^{100}$ as the variance, and an area up to 64.5, or $x = 14.5$, to be evaluated by the normal curve as an approximation of the binomial.)

8. In the fitting of a binomial curve to the scale of sizes, quoted in the preceding chapter, where the sizes were 2, 4, 6, 8, and 10, and the mean was 5.2, the original distribution, and the fitted frequencies as there given, were as follows:

Size (m)	2	4	6	8	10	Σ
Actual	42	99	104	47	8	300
Computed	13	35	34	15	3	100

By chi square determine whether the actual frequencies depart from the theoretical (f_e) more than might be expected on the basis of random sampling. (Note that f_e will be adapted to f by *two* constants, the mean and the total, hence the degrees of freedom are $n = N - 2 = 5 - 2 = 3$.)

APPENDIX

NOTES ON CHAPTER VI

Interpolating in an array.—In order to interpolate quartiles, quintiles, deciles, etc., in an array of ungrouped data, the array may be set up as an irregular tabulation having as its class limits the given items and the means of each two successive items. Unit frequencies are assumed for each class, including the open classes at each end of the tabulation, and N is thus twice the original number of items. The tabulation may be illustrated as follows (array = 25, 30, 33, 37, 45, 55):

L_1	L_2	f	Σ_1	Σ_2
Below	25	1	0	1
25	–27.5	1	1	2
27.5–30		1	2	3
30	–31.5	1	3	4
31.5–33		1	4	5
33	–35	1	5	6
35	–37	1	6	7
37	–41	1	7	8
41	–45	1	8	9
45	–50	1	9	10
50	–55	1	10	11
55 and over		1	11	12

$$N = 12$$

and any percentile, as, for example, the twentieth, or first quintile, may be interpolated by the usual formula, where $FN = 0.20 \times 12 = 2.4$, thus:

$$P = \left(\frac{FN - \Sigma_1}{f}\right)(i) + L_1$$

$$= \left(\frac{2.4 - 2}{1}\right)(2.5) + 27.5 = 28.5$$

That is, the cumulative, 2.4, falls in the class having a lower limit of 27.5, and a class interval of 30–27.5 or 2.5, hence its value is as indicated. Obviously, such an interpolation does not require tabulation except as a means of exposition.

In like manner, the procedure may be reversed to find the relative position of a given magnitude. For example, the magnitude 40 would be located in the class having a lower limit of 37, in which case

$$FN = \left(\frac{P - L_1}{i}\right)(f) + \Sigma_1$$

$$= \left(\frac{40 - 37}{4}\right)(1) + 7 = 7.75$$

and F, the percentile position, is 7.75 ÷ 12 or 0.65.

515

In accordance with the assumptions of this tabulation, the percentile position of each item in the array taken successively is

$$\frac{1}{2N} \quad \frac{3}{2N} \quad \frac{5}{2N}, \dots \frac{2N-1}{2N}$$

where N is the original number of items.

The standard deviation a minimum.—It may be shown that $\Sigma(X-R)^2$ is smallest when R is the mean of the X's.

When the origin is not M, a deviation (d) is $X - M \pm K$, where K is any constant, and

$$d^2 = X^2 + M^2 + K^2 - 2XM \pm 2XK \mp 2MK$$

Summing, noting that $\Sigma X = NM$,

$$\Sigma d^2 = \Sigma X^2 + NM^2 + NK^2 - 2NM^2 \pm 2NMK \mp 2NMK$$

$$= \Sigma X^2 - NM^2 + NK^2$$

But, if R is M, $\Sigma d_M^2 = \Sigma(X-M)^2$, or

$$\Sigma d_M^2 = \Sigma X^2 - 2NM^2 + NM^2 = \Sigma X^2 - NM^2$$

which is smaller than Σd^2, above, by NK^2

Hence, as in the short-cut method of finding σ, Σd^2 from an assumed origin, $M \pm K$ may be "centered" thus:

$$\Sigma d_M^2 = \Sigma d^2 - NK^2$$

or

$$\sigma^2 = \frac{\Sigma d^2}{N} - K^2$$

NOTES ON CHAPTER VII

Probability and curve fitting.—The formulas relating to the logarithmic normal curve are discussed in "The Analysis of Frequency Distributions," by G. R. Davies, *Journal of the American Statistical Association*, December, 1929.

The mathematical aspects of binomial and other types of probability are treated in *Mathematical Statistics*, by H. L. Rietz.

Proof that the normal curve of distribution expressed as $y = e^{-x^2/2}$ has a standard deviation of unity, and that the point of inflection is at $\pm 1\sigma$, may be summarized as follows:

Consider the right half curve from $x = 0$ to $x = \infty$. The area (cf. Peirce, "A Short Table of Integrals," page 63) is $\frac{1}{2}\sqrt{2\pi}$. The standard deviation is $\sigma^2 = \Sigma(x^2 f) \div N$, or, with infinitesimals,

$$\sigma^2 = \frac{2}{\sqrt{2\pi}} \int_0^\infty x^2 e^{-x^2/2}\, dx = \frac{2}{\sqrt{2\pi}} \times \frac{1}{2}\sqrt{2\pi} = 1$$

Hence the standard deviation is unity.

The point on the x scale at which the curve changes from negative to positive curvature (i.e., the point of inflection) is found by equating the second derivative of the curve to zero, and solving for x, thus:

Equation of curve	$y = e^{-x^2/2}$
First derivative	$y' = - e^{-x^2/2} x$
Second derivative	$y'' = e^{-x^2/2}x^2 - e^{-x^2/2} = 0$
Dividing by $e^{-x^2/2}$	$x^2 = 1$, and $x = 1$

Hence the point of inflection is at $x = 1 = \sigma$.

In the usual tables of the normal curve and its area, it is customary so to arrange the ordinates that the total area is unity; that is, the area of the right half is 0.5. As expressed above, however, the area of the half curve is $\frac{1}{2}\sqrt{2\pi}$; or $\sqrt{2\pi} = 2.506628$ for the total curve. It is obvious that the central ordinate of $y = e^{-x^2/2}$ at $x = 0$ is unity, hence to obtain unit area in the total curve it is necessary to divide the ordinates by $\sqrt{2\pi}$. This makes the central ordinate 1 divided by $2.506628 = 0.3989$, and in the curve of unit area $y = (e^{-x^2/2}) \div \sqrt{2\pi}$, where $\sigma = 1$.

Fitting normal curves by graphic means.—Normal curves may be fitted to given data by means of charts such as those shown in Figs. $6 \cdot 2$ and $6 \cdot 3$, as well as by the computations described in Chapter XX. If the data represent an approximately normal distribution, as do those presented in Fig. $6 \cdot 2$, a straight line may be drawn through the points representing the first, second, and third quartiles (Q_1, Q_2, and Q_3), and the ordinates for various values of X may then be identified on the L_2 scale. The desired X value is located on the horizontal scale, after which the point directly above it on the straight line is noted, and the L_2 value is read directly to the left of that point. The value thus determined, after being reduced by the subtraction of 0.50—the remainder being regarded as positive—may be located in the area column of the table of the normal curve, from which the corresponding ordinate may then be read. Ordinate values are recorded in the column designated as z.

Once the z value is known, the actual value of the required ordinate at the given X value is available as

$$Y = z\left(\frac{Ni}{\sigma}\right)$$

The standard deviation of the distribution may be calculated in the usual manner, or it may sometimes be more conveniently determined by graphic means by noting the values at $\sigma = 0$ and at $\sigma = 1$ on the vertical scale of the chart, after which the corresponding points on the X scale may be interpolated by means of the straight line already included in the chart. These two standard deviations are located at cumulatives 50 per cent and 84.13 per cent, respectively.[1] The difference between the two X values thus determined approximates the standard deviation of the distribution. By reference to it, the fitted curve may be measured in terms of a number of its ordinates at various points on the X scale and charted in the usual manner.

---

[1] On the graph paper that is most commonly available, the " less than " cumulative scale is on the right-hand side.

EXAMPLE A1

FITTING A LOGARITHMIC NORMAL CURVE

Data: Assumed skewed distribution,

$$m = 22, \quad 26, \quad 30, \quad 34$$
$$f = 2, \quad 4, \quad 3, \quad 1$$

The quartiles have been computed on the basis of a smoothing process as follows:

$$Q_1 = 24.570; \quad Q_2 = 27.019; \quad Q_3 = 29.675$$

The correction (c) to be added to each X magnitude (class limits, class marks, quartiles, etc.) in order to shift the distribution to a point where it becomes logarithmic as measured by the quartiles, is as follows:

$$c = (Q_2^2 - Q_1 Q_3)/(Q_1 + Q_3 - 2Q_2) = (730.0158 - 729.1265)/0.2079 = 4.278$$

G is $Q_2 + c = 31.297$, and $\log \sigma_r$ is obtained by the following formula:

$$\log \sigma_r = [\log (Q_3 + c) - \log (Q_1 + c)] \times 0.7413 = 0.7413(1.5309 - 1.4601)$$
$$= 0.05246$$

The ordinates of the logarithmic normal curve are fitted for convenience at the upper limits of the classes, though other ordinates may be found by the same process. The upper limits are advanced on the X scale by adding $c = 4.278$. If the corrected quartiles are negative, they may be treated as positive in reverse order. The curve is fitted as follows:

Find $x/\sigma = (\log X - \log G)/\log \sigma_r$; read ordinate (z) and area from table of normal curve of unit area; take $Y = (0.4343 \, Ni/\log \sigma_r)z/X$; Δ_1 as first differences of area from -0.5 to 0.5; and F as Δ_1 times $N = 10$; taking $\log G = 1.4955$, and $\log \sigma_r = 0.05246$. The two extreme F's contain small residuals belonging to more extreme frequencies. The last column shows the deviations of the data from the normal in units of the probable error of sampling. The calculations were carried to more decimal places than are here indicated.

L_2	f	X	Log X	x/σ	z	Y	Area	Δ_1	F	d/PE_s
16	0	20.278	1.3070	—3.5926	0.0006	0.0101	—0.4998	0.0002	0.002	0.07
20	0	24.278	1.3852	—2.1024	0.0438	0.5970	—0.4822	0.0176	0.176	0.63
24	2	28.278	1.4514	—0.8397	0.2804	3.2837	—0.2995	0.1828	1.828	0.21
28	4	32.278	1.5089	0.2556	0.3861	3.9611	0.1009	0.4003	4.003	0.00
32	3	36.278	1.5596	1.2226	0.1889	1.7245	0.3893	0.2884	2.884	0.12
36	1	40.278	1.6051	2.0886	0.0450	0.3703	0.4816	0.0924	0.924	0.12
40	0	44.278	1.6462	2.8725	0.0064	0.0482	0.4980	0.0163	0.163	0.60
44	0	48.278	1.6837	3.5884	0.0006	0.0043	0.4998	0.0019	0.020	0.21

To find the mode (Mo) and the ordinate at the mode (Y_{Mo}):
Let a = antilog $[(\log \sigma_r)^2/0.4343]$
Then $Mo = G/a = 31.297/1.0147 = 30.844$
and $Mo - c = 26.566$ on original X scale.
$$Y_{Mo} = (0.17326 \, ni/\log \sigma_r) \, a^{1/2}/G$$
$$= 132.108 \times 1.00732 \div 31.2968 = 4.2520.$$

Y may be computed, without reading z from a table of the normal curve, by the formulas (where L_s is log σ_r),

$$Y = G \div X(2\pi)^{1/2} e^{1/2 \left[\left(\log \frac{X}{G} \right) + L_s \right]^2}$$

or, as adjusted for purposes of calculation (area as given),

$$Y = \frac{0.17326 \; ni/L_s}{X \text{ antilog } [(0.21715/L_s^2)(\log X - \log G)^2]}$$

If the distribution is positively skewed, its cumulatives, expressed in terms of log L_2, should be plotted on the probability chart. The logarithmic quartiles may then be interpolated and their antilogarithms regarded as the quartiles of the distribution. The quartiles may then be used to provide a correction for skewness, a value that may be defined as

$$c = \frac{Q_2^2 - Q_1 Q_3}{Q_1 + Q_3 - 2Q_2}$$

If the numerator of this figure is zero or near zero, the distribution may be regarded as logarithmic normal without correction. Otherwise, c may be added to each item in the X series (the limits as well as the class marks), and the chart redrawn (each cumulative against the logarithm of its revised upper limit). In the new graph, the quartiles should fall in approximately a straight line. If they fail to do so appreciably, a second correction may be made in the same manner. If the corrected quartiles are negative, they may be treated as positive in reverse order.

When the corrected chart permits the joining of the quartiles by a straight line, any required ordinate may be found by reading from the log X scale to that straight line, thence directly to the left to the cumulative scale, and from that value (utilizing the table of ordinates and areas), as above, to the value of the normal ordinate, z. The required ordinate is (X is antilog of log X)

$$Y = \frac{z}{X} \times \frac{0.4343ni}{\log \sigma_r}$$

The term, log σ_r, may be found on the log L_2-scale just as σ was found in the normal distribution described above. After enough ordinates have been determined, the log normal curve may be plotted against the original X's; i.e., $X - c$. A calculation paralleling this graphic procedure is detailed in Example A1.

In some distributions having negative skewness the same technique may be employed if the X scale is reversed to read from an origin at some point above the distribution. For example, a distribution of percentage grades may be taken in reverse with 100 as the origin (0).

NOTES ON CHAPTER IX

INDEX NUMBERS

Fisher's formula.—The analysis of Fisher's ideal index number may be indicated as follows:

$$Q_b \times P_r = V: \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

and similarly

$$Q_r \times P_b = V$$

Hence,

$$P_i Q_i = (P_b P_r)^{1/2}(Q_b Q_r)^{1/2} = (P_b Q_r)^{1/2}(P_r Q_b)^{1/2} = V^{1/2}V^{1/2} = V$$

Or the same evidence of consistency may be adduced by writing the formulas in full, thus:

$$Q_i = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}}$$

$$P_i = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

If these two formulas are multiplied, we have:

$$P_i Q_i = \sqrt{\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}} = V$$

NOTES ON CHAPTERS X AND XI

It is sometimes desirable to change the origin of the X scale after a trend equation has been calculated. Suppose that it seems desirable for purposes of comparison with other data to express the trend equation so that it will be applicable to a new X scale ($\bar{X}$) having a selected origin R. The constants ($\bar{a}$ and $\bar{b}$) of the equation applicable to $\bar{X}$, in terms of the a and b as calculated, are

$$\bar{a} = a + bR$$
$$\bar{b} = b$$

A parabolic trend equation may similarly be adjusted to a new origin by the equations

$$\bar{a} = a + bR + cR^2$$
$$\bar{b} = b + 2cR$$
$$\bar{c} = c$$

If more radical changes in the scaling are required, decoding equations (see page 546) may be employed.

TABLE A1

CONSTANTS OF PARABOLAS, BY WEIGHTS

The following weights may be applied to successive items (A, B, C, etc.) to find the constants of $Y = a + bx + cx^2$; origin at mid-point. Multiply each item by its weight, total, and divide as indicated. Note that, for b and c, the divisor is not the sum of the weights.

		Items of data, at successive unit intervals															Divisor
N	p	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
3	a	0	1	0													1
	b	−1	0	1													2
	c	1	−2	1													2
4	a	−1	9	9	−1												16
	b	−3	−1	1	3												10
	c	1	−1	−1	1												4
5	a	−3	12	17	12	−3											35
	b	−2	−1	0	1	2											10
	c	2	−1	−2	−1	2											14
6	a	−3	7	12	12	7	−3										32
	b	−5	−3	−1	1	3	5										35
	c	5	−1	−4	−4	−1	5										56
7	a	−2	3	6	7	6	3	−2									21
	b	−3	−2	−1	0	1	2	3									28
	c	5	0	−3	−4	−3	0	5									84
8	a	−3	3	7	9	9	7	3	−3								32
	b	−7	−5	−3	−1	1	3	5	7								84
	c	7	1	−3	−5	−5	−3	1	7								168
9	a	−21	14	39	54	59	54	39	14	−21							231
	b	−4	−3	−2	−1	0	1	2	3	4							60
	c	28	7	−8	−17	−20	−17	−8	7	28							924
10	a	−14	6	21	31	36	36	31	21	6	−14						160
	b	−9	−7	−5	−3	−1	1	3	5	7	9						165
	c	6	2	−1	−3	−4	−4	−3	−1	2	6						264
11	a	−36	9	44	69	84	89	84	69	44	9	−36					429
	b	−5	−4	−3	−2	−1	0	1	2	3	4	5					110
	c	15	6	−1	−6	−9	−10	−9	−6	−1	6	15					858
12	a	−9	1	9	15	19	21	21	19	15	9	1	−9				112
	b	−11	−9	−7	−5	−3	−1	1	3	5	7	9	11				286
	c	55	25	1	−17	−29	−35	−35	−29	−17	1	25	55				4004
13	a	−11	0	9	16	21	24	25	24	21	16	9	0	−11			143
	b	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6			182
	c	22	11	2	−5	−10	−13	−14	−13	−10	−5	2	11	22			2002
14	a	−33	−3	22	42	57	67	72	72	67	57	42	22	−3	−33		448
	b	−13	−11	−9	−7	−5	−3	−1	1	3	5	7	9	11	13		455
	c	13	7	2	−2	−5	−7	−8	−8	−7	−5	−2	2	7	13		1456
15	a	−78	−13	42	87	122	147	162	167	162	147	122	87	42	−13	−78	1105
	b	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	280
	c	91	52	19	−8	−29	−44	−53	−56	−53	−44	−29	−8	19	52	91	12376

TABLE A2

FITTING PARABOLA TRENDS BY WEIGHTS

The following weights (w) are applied to successive items (A, B, C, etc.) to find the parabola trend at the designated x. The data are assumed to be at unit intervals, origin at center; i.e., $\Sigma x = 0$. Multiply each item of the data by the appropriate weight in the column below, total, and divide by Σ weights. Two extrapolations of positive x's are given. To obtain trend items for a negative x, use the weights in reverse order. To interpolate an item ($N > 3$) enter it as 0, deduct its w from Σw, and find its T.

N	x	\multicolumn{13}{Items of data, at successive unit intervals}												Σ weights
		A	B	C	D	E	F	G	H	I	J	K	L	M	
3	0	0	1	0											1
	1	0	0	1											1
	2	1	-3	3											1
4	0.5	-3	9	11	3										20
	1.5	1	-3	3	19										20
	2.5	3	-5	-3	9										4
	3.5	39	-57	-43	81										20
5	0	-3	12	17	12	-3									35
	1	-5	6	12	13	9									35
	2	3	-5	-3	9	31									35
	3	3	-3	-4	0	9									5
	4	7	-6	-9	-2	15									5
6	0.5	-5	6	12	13	9	0								35
	1.5	-15	7	24	36	43	45								140
	2.5	3	-3	-4	0	9	23								28
	3.5	5	-3	-6	-4	3	15								10
	4.5	75	-37	-84	-66	17	165								70
7	0	-2	3	6	7	6	3	-2							21
	1	-2	1	3	4	4	3	1							14
	2	-1	0	1	2	3	4	5							14
	3	5	-3	-6	-4	3	15	32							42
	4	3	-1	-3	-3	-1	3	9							7
	5	12	-3	-11	-12	-6	7	27							14
8	0.5	-21	11	33	45	47	39	21	-7						168
	1.5	-7	1	7	11	13	13	11	7						56
	2.5	-7	-3	3	11	21	33	47	63						168
	3.5	3	-1	-3	-3	-1	3	9	17						24
	4.5	21	-3	-17	-21	-15	1	27	63						56
	5.5	119	-9	-87	-115	-93	-21	101	273						168

Group	x															Total
9	0	-21	14	39	54	59	54	39	14	-21						231
	1	-294	42	293	459	540	536	447	273	14						2,310
	2	-238	-21	156	293	390	447	464	441	378						2,310
	3	-6	-7	-3	6	20	39	63	92	126						330
	4	21	-3	-17	-21	-15	1	27	63	109						165
	5	14	0	-9	-13	-12	-6	5	21	42						42
	6	126	7	-72	-111	-110	-69	12	133	294						210
	6.5	-18	3	19	30	36	37	33	24	10						165
10	1.5	-26	-2	17	31	40	44	43	37	26	-9					220
	2.5	-54	-14	21	51	76	96	111	121	126	10					660
	3.5	0	-6	-7	-3	6	20	39	63	92	126					330
	4.5	14	0	-9	-13	-12	-6	5	21	42	126					110
	5.5	18	2	-9	-15	-16	-12	-3	11	30	68					60
	6.5	342	54	-149	-267	-300	-248	-111	111	418	54					660
11	0	-36	9	44	69	84	89	84	69	44	46	-36				429
	1	-80	-6	53	97	126	140	139	123	92	87	-15				715
	2	-75	-17	31	69	97	115	123	121	109	414	55				715
	3	-135	-54	22	93	159	220	276	327	373	199	450				2,145
	4	10	-9	-18	-17	-6	15	46	87	138	54	270				715
	5	18	2	-9	-15	-16	-12	-3	11	30	81	83				143
	6	45	9	-17	-33	-39	-35	-21	-3	37	34	135				165
12	0.5	25	6	-8	-17	-21	-20	-14	-3	13	53	15	-33			55
	1.5	-55	-3	39	71	93	105	107	99	81	471	281	33			572
	2.5	-429	-97	177	393	551	651	693	677	603	573	507	407			4,004
	3.5	-363	-123	83	255	393	497	567	603	605	677	783	891			4,004
	4.5	-187	-99	-9	83	177	273	371	471	573	783	1,109	1,485			4,004
	5.5	99	-25	-99	-123	-97	-21	105	281	507	81	135	199			4,004
	6.5	45	9	-17	-33	-39	-35	-21	3	37	11	21	33			364
	7.5	11	3	-3	-7	-9	-9	-7	-3	3	159	341	561			44
13	0	231	71	-51	-135	-181	-189	-159	-91	15	16	9	0	-11		572
	1	-11	0	9	16	21	24	25	24	21	126	86	33	-33		143
	2	-99	-22	42	93	131	156	168	167	153	135	111	77	33		1,001
	3	-99	-33	23	69	105	131	147	153	149	139	138	132	121		1,001
	4	-77	-33	6	40	69	93	112	126	135	138	167	198	231		1,001
	5	-33	-22	-9	6	23	42	63	86	111	12	18	25	33		1,001
	6	3	0	-2	-3	-3	-2	0	3	7	11	21	33	47		91
	7	11	3	-3	-7	-9	-9	-7	-3	3	15	38	66	99		91
	8	363	132	-49	-180	-261	-292	-273	-204	-85	84	303	572	891		1,001

TABLE A2—*Continued*

FITTING PARABOLA TRENDS BY WEIGHTS

N	x				Items of data, at successive unit intervals												Σ Weights
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
14	0.5	−39	−8	18	39	55	66	72	73	69	60	46	27	3	−26		455
	1.5	−351	−117	82	246	375	469	528	552	541	495	414	298	147	−39		3,640
	2.5	−65	−27	6	34	57	75	88	96	99	97	90	78	61	39		728
	3.5	−117	−59	−6	42	85	123	156	184	207	225	238	246	249	247		1,820
	4.5	−39	−33	−22	−6	15	41	72	108	149	195	246	302	363	429		1,820
	5.5	143	21	−66	−118	−135	−117	−64	24	147	305	498	726	989	1,287		3,640
	6.5	33	11	−6	−18	−25	−27	−24	−16	−3	15	38	66	99	137		280
	7.5	39	15	−4	−18	−27	−31	−30	−24	−13	3	24	50	81	117		182
	8.5	299	123	−18	−124	−195	−231	−232	−198	−129	−25	114	288	497	741		910
15	0	−78	−13	42	87	122	147	162	167	162	147	122	87	42	−13	−78	1,105
	1	−2,730	−897	671	1,974	3,012	3,785	4,293	4,536	4,514	4,227	3,675	2,858	1,776	429	−1,183	30,940
	2	−2,821	−1,170	261	1,472	2,463	3,234	3,785	4,116	4,227	4,118	3,789	3,240	2,471	1,482	273	30,940
	3	−2,457	−1,183	−54	930	1,769	2,463	3,012	3,416	3,675	3,789	3,758	3,582	3,261	2,795	2,184	30,940
	4	−819	−468	−137	174	465	736	987	1,218	1,429	1,620	1,791	1,942	2,073	2,184	2,275	15,470
	5	−364	−429	−399	−274	−54	261	671	1,176	1,776	2,471	3,261	4,146	5,126	6,201	7,371	30,940
	6	105	26	−33	−72	−91	−90	−69	−28	33	114	215	336	477	638	819	2,380
	7	39	15	−4	−18	−27	−31	−30	−24	−13	3	24	50	81	117	158	340
	8	91	39	−3	−35	−57	−69	−71	−63	−45	−17	21	69	127	195	273	455
	9	546	247	3	−186	−320	−399	−423	−392	−306	−165	31	282	588	949	1,365	1,820

Interpolation formulas.—If equispaced ordinates, A, B, and C, are given, at unit intervals of $-x$, 0, and x, respectively, parabolic interpolations may be readily made. By Table A1 (origin at B),

$$Y = B + \tfrac{1}{2}(C - A)x + \tfrac{1}{2}(A - 2B + C)x^2$$

which simplifies to

$$Y = (1 - x^2)B + \tfrac{1}{2}x(1 + x)C - \tfrac{1}{2}x(1 - x)A$$

If $x = \tfrac{1}{2}$ is substituted in this equation, a parabolic interpolation mid-way between B and C is obtained as (origin at B),

$$Y_{\frac{1}{2}} = (6B + 3C - A) \div 8$$

and the interpolation at $x = -1$ may be obtained by interchanging A and C, that is (origin at B),

$$Y_{-\frac{1}{2}} = (6B + 3A - C) \div 8$$

Similarly (origin at B),

$$Y_{\frac{1}{3}} = (8B + 2C - A) \div 9$$

$$Y_{\frac{2}{3}} = (5B + 5C - A) \div 9$$

and

$$Y_{\frac{1}{4}} = (30B + 5C - 3A) \div 32$$

$$Y_{\frac{1}{2}} = (6B + 3C - A) \div 8$$

$$Y_{\frac{3}{4}} = (14B + 21C - 3A) \div 32$$

and negative x interpolations may be made by interchanging A and C, as before.

If A, B, and C are annual data, and quarterly interpolations are required, they may be made at $x = \tfrac{1}{8}$ (third quarter); $x = \tfrac{3}{8}$; $x = \tfrac{5}{8}$; and $x = \tfrac{7}{8}$, thus (origin at B):

$$Y_{\frac{1}{8}} = (126B + 9C - 7A) \div 128$$

$$Y_{\frac{3}{8}} = (110B + 33C - 15A) \div 128$$

$$Y_{\frac{5}{8}} = (78B + 65C - 15A) \div 128$$

$$Y_{\frac{7}{8}} = (30B + 105C - 7A) \div 128$$

and Y's at negative x's may be obtained as before.

Smoothed interpolations between successive equispaced Y's may be made as weighted averages of parabolic interpolations. Thus, in the points A, B, C, and D, interpolations between B and C may be made by averaging positive interpolations for A, B, and C (origin at B) and negative for B, C, and D (origin at C), weights being $1 - |x|$. Thus, three smoothed interpolations between B and C, in the series A, B, C, and D, would be, successively,

x	Y (origin at B)	x	Y (origin at C)
$\tfrac{1}{4}$	$(30B + 5C - 3A) \div 32$	$-\tfrac{3}{4}$	$(14C + 21B - 3D) \div 32$
$\tfrac{1}{2}$	$(6B + 3C - A) \div 8$	$-\tfrac{1}{2}$	$(6C + 3B - D) \div 8$
$\tfrac{3}{4}$	$(14B + 21C - 3A) \div 32$	$-\tfrac{1}{4}$	$(30C + 5B - 3D) \div 32$

Average (weights $= 1 - |x|$)

$$(111B + 29C - 3D - 9A) \div 128$$

$$(9B + 9C - D - A) \div 16$$

$$(29B + 111C - 9D - 3A) \div 128$$

If the four points A, B, C, and D, only, are involved, the smoothing may be completed by parabolic interpolation between A and B, utilizing C, and between D and C, utilizing B. But if further points are included, the items preceding the space to be interpolated may be taken as A and B, and the next two following as C and D.

Similar interpolations at x, as $\frac{1}{8}$, $\frac{3}{8}$, $\frac{5}{8}$, or $\frac{7}{8}$, suitable for interpolations of quarters in annual data, or, with the weighted averages above, of interpolations by eighths, are successively as follows:

x	Y (origin at B)
$\frac{1}{8}$	$(987B + 93C - 7D - 49A) \div 1,024$
$\frac{3}{8}$	$(745B + 399C - 45D - 75A) \div 1,024$
$\frac{5}{8}$	$(399B + 745C - 75D - 45A) \div 1,024$
$\frac{7}{8}$	$(93B + 987C - 49B - 7A) \div 1,024$

Other interpolation formulas may be made as required.

Aids in calculating trend equations.—The equations for the constants of straight line and parabolic trends may be more readily solved by means of the accompanying table (A3) of certain designated functions of N and x, assuming centered time scale (x) having unit intervals so that $\Sigma x = 0$ (e.g., -1; 0; 1; or -1.5; -0.5; 0.5; 1.5).

<p align="center">TABLE A3</p>

<p align="center">SUMMATIONS FOR FITTING PARABOLAS</p>

N	Σx^2	$(N\Sigma x^4 - \Sigma x^2 \Sigma x^2)$	N	Σx^2	$(N\Sigma x^4 - \Sigma x^2 \Sigma x^2)$
2	0.5	0	22	885.5	623,392
3	2	2	23	1,012	814,660
4	5	16	24	1,150	1,052,480
5	10	70	25	1,300	1,345,500
6	17.5	224	26	1,462.5	1,703,520
7	28	588	27	1,638	2,137,590
8	42	1,344	28	1,827	2,660,112
9	60	2,772	29	2,030	3,284,946
10	82.5	5,280	30	2,247.5	4,027,520
11	110	9,438	31	2,480	4,904,944
12	143	16,016	32	2,728	5,936,123
13	182	26,026	33	2,992	7,141,904
14	227.5	40,768	34	3,272.5	8,545,152
15	280	61,880	35	3,570	10,170,930
16	340	91,392	36	3,885	12,046,608
17	408	131,784	37	4,218	14,202,006
18	484.5	186,048	38	4,569.5	16,669,536
19	570	257,754	39	4,940	19,484,348
20	665	351,120	40	5,330	22,684,480
21	770	471,086			

Trend adjustments.—Trends fitted by indirect methods applied to the logarithms, reciprocals, etc., will sometimes fail to provide a satisfactory fit because of distortions arising from the changing of the scale, or for similar reasons. They may however,

be used as one element in a least-squares adjustment, in which a constant and slope are the other two elements. This method, however, is tedious and may be approximated as follows. Compute the equation of a straight-line trend ($T = a + bx$) for both the data (Y) and the approximate trend (T_a), taking the time scale centered. From the equation for Y, subtract the corresponding equation for T_a. Compute the straight-line trend indicated by the equation thus found. This is the correction trend (T_c), which when added to T_a will give a final trend (T) which has the same sum and the same slope as the data. Ordinarily, this adjustment will improve the fit of the calculated trend. The method is illustrated in the accompanying example.

EXAMPLE A2

CORRECTING AN APPROXIMATE TREND (T_a)

(Assumed to have been calculated by an indirect, inexact method)

x	Y	T_a	T_c	T
-2	20	10	0	10
-1	10	28	2	30
0	40	36	4	40
1	60	34	6	40
2	20	22	8	30
	$5\overline{)150}$	$5\overline{)130}$		
	$a = 30$	26		$T_c = (30 - 26) + (5 - 3)x = 4 + 2x$
	$b = 5$	3		$T = T_a + T_c$

The method may be extended to include a parabolic correction.

The direct least-squares trend.—Under certain conditions, illustrated below, trends may be fitted by least squares measured perpendicularly to the trend. Such lines are described as fitted by direct least squares. The method of fitting is somewhat more complex than the usual method although it makes use of the same quantities, Σxy, Σx^2, and Σy^2. The equation is, as before,

$$T = a + bx$$

and a also is found as before, that is, it is

$$a = \frac{\Sigma Y}{N}$$

or, if the data are centered, it is zero. The equation for b being the solution of a quadratic, has alternate values

$$b = \frac{1 + \sqrt{w^2 + 1}}{-w} \quad \text{or} \quad \frac{w}{1 + \sqrt{w^2 + 1}}$$

where

$$w = \frac{2\Sigma xy}{\Sigma x^2 - \Sigma y^2}$$

The required value of b has the sign of Σxy.

This type of straight-line trend has been used in the calculation of the elasticity of the market, where elasticity is measured by the coefficient $\eta = 1/b$. The usual technique is as follows:

1. Tabulate the necessary data, a sequence of production and deflated prices, usually year-by-year for the given product in question.

2. Compute the link relatives of each of these series, that is, express each item (except the first) in percentages of the preceding item taken as 100 per cent.

3. Write the logarithms of these link relatives. These two series of logarithms are regarded as X and Y, respectively, of the calculations (X refers to quantities and Y to prices).

4. Calculate Σxy, Σy^2, and Σx^2 either by actually centering the data or by the use of the usual centering formulas, or both.

5. Apply the equations mentioned above to obtain w and the two values of b. The appropriate value of b may be determined by the sign of Σxy; if this is positive, the positive b is chosen; if negative the negative b is chosen. When zero or infinity is obtained in the fraction w, the result may readily be interpreted by plotting. Example A3 is a very simplified illustration.

Other equations.—There are many methods of fitting parabolas, particularly as applied to regular time series with the time scale centered at the average date. One of these is represented by the following equations, which are merely the equations employed in Chapter XI transformed so as to utilize the quantities, as obtained from the data: N, N^2, ΣY, ΣxY, Σx^2Y.
The equations are:

$$a = \frac{3(3N^2 - 7)\Sigma Y - 60\Sigma x^2Y}{4N(N^2 - 4)}$$

$$b = \frac{12\Sigma xY}{N(N^2 - 1)}$$

$$c = \frac{180\Sigma x^2Y - 15(N^2 - 1)\Sigma Y}{N(N^2 - 1)(N^2 - 4)}$$

The principal advantage of these equations is that tables may readily be constructed giving the functions of N, i.e., N^2, $N^2 - 1$, $N^2 - 4$, etc., thus making readily available all the factors in the equation with the exception of the three summations containing Y.

Equations for fitting a parabola by the method of grouped data are given and illustrated in a succeeding paragraph. Equations suitable for use with the method of selected points may also be derived. It will be recalled that in this method the data are charted, and three points (P_1, P_2 and P_3) estimated to lie on the trend, on equidistant ordinates t intervals apart, and located near the beginning, the middle, and the end of the series, respectively, are read from the chart. The constants of the equation are (origin at P_2):

$$a = P_2$$

$$b = \frac{P_3 - P_1}{2t}$$

$$c = \frac{P_1 + P_3 - 2P_2}{2t^2}$$

EXAMPLE A3

ELASTICITY OF A MARKET

Data: Assumed production and deflated price of crop A in a total market.

Years	Data		Link relative		Logarithms of P and Q		Deviations of Y and X				
	P	Q	P	Q	Y	X	y	x	y^2	x^2	xy
1910	$1.00	240									
1911	1.26	223	126	93	2.10	1.97	0.10	−0.03	0.0100	0.0009	−0.0030
1912	1.15	230	91	103	1.96	2.01	−0.04	0.01	0.0016	0.0001	−0.0004
1913	1.15	230	100	100	2.00	2.00	0.00	0.00	0.0000	0.0000	0.0000
1914	1.21	216	105	94	2.02	1.97	0.02	−0.03	0.0004	0.0009	−0.0006
1915	1.00	244	83	113	1.92	2.05	−0.08	0.05	0.0064	0.0025	−0.0040
					5)10.00	10.00	0.00	0.00	0.0184	0.0044	−0.0080
					2.00	2.00					

Check:

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = \begin{array}{r} 20.0044 \\ -20.0000 \\ \hline 0.0044 \end{array} \quad \text{(use 44)}$$

$$\Sigma y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = \begin{array}{r} 20.0184 \\ -20.0000 \\ \hline 0.0184 \end{array} \quad \text{(use 184)}$$

$$\Sigma xy = \Sigma XY - \frac{\Sigma X \Sigma Y}{N} = \begin{array}{r} 19.9920 \\ -20.0000 \\ \hline -0.0080 \end{array} \quad \text{(use −80)}$$

$$w = \frac{2\Sigma xy}{\Sigma x^2 - \Sigma y^2} = \frac{-160}{44 - 184} = 1.143$$

$$b = \frac{1 \pm \sqrt{w^2 + 1}}{-w} = \frac{1 + \sqrt{2.306}}{-1.143} = \frac{2.519}{-1.143} = -2.204$$

$$\text{Elasticity} = \frac{1}{b} = \frac{1}{-2.204} = -0.45$$

The results roughly approximate the wheat and corn market.

For similar equations applicable to a cubic, see Davies and Crowder, *Methods of Statistical Analysis*, Chapter VI.

The parabola by grouped data.—The fitting of a parabola may be somewhat abbreviated by applying the criterion of grouped data, which sets up the standard that, when the series is regularly spaced, and when N is divisible by 3, the sum of the trend items in the first group ($m = N/3$ items) should equal the sum of the data; and that this equality should be true in the second and third group of m items, respectively. This method should not be confused with least-squares fittings to averages of the data by decades, or other periods. The method is illustrated in Example A4.

<div align="center">

EXAMPLE A4

A PARABOLA FITTED BY THE GROUPED DATA METHOD

</div>

Data: Assumed production, 1901–1906.

X	Y	S	x	x^2	a	$+$	bx	$+$	cx^2		T
1901	70	S_1	-2.5	6.25	100		-10		-25	$=$	65
1902	80	150	-1.5	2.25	100		-6		-9	$=$	85
1903	98	S_2	-0.5	0.25	100		-2		-1	$=$	97
1904	100	198	0.5	0.25	100		$+2$		-1	$=$	101
1905	95	S_3	1.5	2.25	100		$+6$		-9	$=$	97
1906	87	182	2.5	6.25	100		$+10$		-25	$=$	85
		530		17.50							530

$$b = (S_3 - S_1) \div 2m^2 = (182 - 150) \div (2 \times 4) = 4$$
$$c = (S_1 + S_3 - 2S_2) \div 2m^3 = (150 + 182 - 396) \div (2 \times 8) = -4$$
$$a = (S_2 - c\Sigma x_2^2) \div m = [198 - (-4)(0.5)] \div 2 = 100$$

It should be noted that S_1, S_2, and S_3 are the sums of the first, second, and third group of m items, respectively. Also the expression Σx_2^2 means the sum of the squares of the x's (centered) in the mid-group.

The normal equations of parabolas.—The derivation of the equations used with the straight-line and parabola trends may be briefly suggested. It will be seen that, in the general equation, $T = a + bX + cX^2$, a determines an origin height; bX, a slope; and cX^2, a constant curvature. To equate these measures in the data (Y) and the trend (T), we assume that $\Sigma Y = \Sigma T$; $\Sigma XY = \Sigma XT$; and $\Sigma X^2 Y = \Sigma X^2 T$. If $a + bX + cX^2$ is substituted for T in each of these equalities, the so-called normal equations are obtained:

$$\Sigma Y = Na + b\Sigma X + c\Sigma X^2$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$\Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

If X is unit spaced and centered, these equations are readily simplified.

That the parabola trend fitted to data by the least-squares formulas has the least possible standard deviation ($\Sigma \overline{Y - T}^2$ a minimum) of any curve of its type may be proved by the calculus method of derivation, from the equation, where $d = Y - T$

$$\Sigma d^2 = \Sigma(Y - T)^2 = \Sigma(Y - a - bX - cX^2)^2$$

The first derivatives with respect to a, b, and c, at the minimum of Σd^2, are $du^n/da = nu^{n-1}du/da$, where $u = \Sigma(Y - a - bX - cX^2)$, etc., or

$$2\Sigma(Y - a - bX - cX^2)(-1) = 0$$

$$2\Sigma(Y - a - bX - cX^2)(-X) = 0$$

$$2\Sigma(Y - a - bX - cX^2)(-X^2) = 0$$

which reduce to the three normal equations, respectively. The same proof may be applied to a straight line or an exponential of any degree.

The modified reciprocal trend.—A substitute for the modified geometric trend as applied to a series of data increasing at a declining rate so as to approach virtually a horizontal asymptote, or to their logarithms, is the modified reciprocal,[1] the chief element of which consists of the reciprocals of the adjusted X scale. Its equation is

$$T = a + \frac{b}{x + c}$$

It may be fitted by plotting the data and estimating three trend points, P_1, P_2, and P_3, on equidistant ordinates near the beginning, middle, and end of the series, respectively. These ordinates are designated as $x = -t$, $x = 0$, and $x = t$, respectively, where t is the number of time units between selected points. Then

$$c = \frac{t(P_3 - P_1)}{2P_2 - (P_1 + P_3)}$$

$$a = \frac{P_2(P_1 + P_3) - 2P_1P_3}{2P_2 - (P_1 + P_2)}$$

$$b = c(P_2 - a)$$

For example, suppose that the selected points are

$$T = P_1 = 10; \quad P_2 = 20; \quad P_3 = 25 \text{ when } t = 4$$

Then

$$x = -4; \quad 0; \quad 4$$

and

$$c = \frac{4(25 - 10)}{2 \times 20 - (10 + 25)} = 12$$

$$a = \frac{20(10 + 25) - 2 \times 10 \times 25}{2 \times 20 - (10 + 25)} = 40$$

$$b = 12(20 - 40) = -240$$

The trend passes through the three selected points, as follows:

$$T = a + \frac{b}{(x + c)} = 40 - \frac{240}{x + 12}$$

[1] See Norris O. Johnson, "A Trend Line for Growth Series," *Journal of the American Statistical Association*, 30 (192), December, 1935, p. 717.

$$\text{at } x = -4: T = 40 - \frac{240}{12 - 4} = 10$$

$$\text{at } x = 0 : T = 40 - \frac{240}{0 + 12} = 20$$

$$\text{at } x = 4 : T = 40 - \frac{240}{4 + 12} = 25$$

And intermediate points on the trend may similarly be found.   For example,

$$\text{at } x = 2 : T = 40 - \frac{240}{2 + 12} = 22.86$$

As a rule this trend is preferably fitted to the logarithms of P_1, P_2, and P_3, and the antilogarithms of the trend thus found are taken as the final trend.

The method may be improved by first estimating c as before and then using the method of least squares. As thus applied, the series $\dfrac{1}{x + c}$ is taken as the x series of the normal equations. The closeness of fit may be still further improved by adding another term, dx.

The modified geometric trend.—The modified geometric trend may be conveniently fitted by the method of grouped data as follows: Assume that the number of items is divisible by 3 or can readily be made so by dropping one or two of the initial items. The series to be fitted is arranged in three consecutive groups of $m = N/3$ items each. If it does not appear suitable to drop one or two items so that N is divisible by 3, the method can be adjusted to fractional items, or each item in the original data may be repeated three times, representing the first, second, and third four-month periods in each year and the trend later taken as of the central four-months. Or one or two additional items may be extrapolated in such a way that they will practically fall upon the trend. This may. be done by parabolic extrapolation, employing the weights given on pages 522–524.   The method of fitting the trend is illustrated in Example A4′. The data (Y) are arranged in three subtotals $(S_1, S_2, \text{and } S_3)$, and the first differences $(d_1 \text{ and } d_2)$ are taken. The equations of the constants as given below are then applied. The general equation is

$$T = a + bc^x; \quad m = \frac{N}{3}$$

<div align="center">

EXAMPLE A4'

THE MODIFIED GEOMETRIC TREND; METHOD OF GROUPED DATA

</div>

Data: Assumed production indexes.

Years	x	Y	S	d	$100 + 5 \times 2^x = T$		
1900	0	104			100	5	105
1901	1	111	$S_1 = 215$		100	10	110
1902	2	122		$d_1 = 45$	100	20	120
1903	3	138	$S_2 = 260$		100	40	140
1904	4	184		$d_2 = 180$	100	80	180
1905	5	256	$S_3 = 440$		100	160	260

$$c^m = \frac{d_2}{d_1}; \quad c^2 = \frac{180}{45} = 4; \quad c = 2$$

$$b = \frac{d_1(c-1)}{(c^m-1)^2} = \frac{45 \times 1}{3^2} = 5$$

$$ma = S_1 - \frac{d_1}{c^m-1}; \quad 2a = 215 - \frac{45}{3} = 200; \quad a = 100$$

When the modified geometric trend is fitted by the method of selected points to annual or other regular data (P_1, P_2, and P_3 are trend points estimated on a chart at t years apart) the following equations may be utilized (origin at P_1):

$$c^t = \frac{P_3 - P_2}{P_2 - P_1}$$

$$b = \frac{P_2 - P_1}{c^t - 1}$$

$$a = P_1 - b$$

These equations may obviously be adapted to the Pearl-Reed and Gompertz curves by fitting the modified geometric trend to the reciprocals or logarithms of the selected points, respectively, and reversing the trend thus found by utilizing its reciprocals or logarithms.

Sums of powers.—In many trend and correlation problems, formulas are used which require the powers of a series of numbers, often centered as deviations from the mean. In such cases the formulas of Table A4 will prove convenient.

TABLE A4

SUMMATION FORMULAS

I. For any variate, X, where $X - M$ is written as x:

(1) $\Sigma x = 0$

(2) $\Sigma x^2 = \Sigma X^2 - \dfrac{(\Sigma X)^2}{N}$

(3) $\Sigma x^3 = \Sigma X^3 - \dfrac{3\Sigma X^2 \Sigma X}{N} + 2\dfrac{(\Sigma X)^3}{N^2}$

(4) $\Sigma x^4 = \Sigma X^4 - \dfrac{4\Sigma X^3 \Sigma X}{N} + 6\Sigma X^2 \dfrac{(\Sigma X)^2}{N^2} - \dfrac{3(\Sigma X)^4}{N^3}$

II. Sums of items in certain designated series:

(N is number of items in series; increasing series begin as indicated, and may be carried to N items; decreasing-increasing and increasing-decreasing series are symmetrical and may be carried higher.)

No.		SERIES
(1)	$0 + 1 + 2 + \ldots\ldots\ldots\ldots\ldots\ldots$	$= N(N-1)/2$
(2)	$0^2 + 1^2 + 2^2 + \ldots\ldots\ldots\ldots\ldots$	$= N(N-1)(2N-1)/6$
(3)	$0^3 + 1^3 + 2^3 + \ldots\ldots\ldots\ldots\ldots$	$= N^2(N-1)^2/4$
(4)	$1 + 2 + 3 + \ldots\ldots\ldots\ldots\ldots\ldots$	$= N(N+1)/2$
(5)	$1^2 + 2^2 + 3^2 + \ldots\ldots\ldots\ldots\ldots$	$= N(N+1)(2N+1)/6$
(6)	$1^3 + 2^3 + 3^3 + \ldots\ldots\ldots\ldots\ldots$	$= N^2(N+1)^2/4$
(7)	$0 + 2 + 4 + \ldots\ldots\ldots\ldots\ldots\ldots$	$= N(N-1)$
(8)	$0^2 + 2^2 + 4^2 + \ldots\ldots\ldots\ldots\ldots$	$= 2N(N-1)(2N-1)/3$
(9)	$0^3 + 2^3 + 4^3 + \ldots\ldots\ldots\ldots\ldots$	$= 2N^2(N-1)^2$
(10)	$1 + 3 + 5 + \ldots\ldots\ldots\ldots\ldots\ldots$	$= N^2$
(11)	$1^2 + 3^2 + 5^2 + \ldots\ldots\ldots\ldots\ldots$	$= N(4N^2-1)/3$
(12)	$1^3 + 3^3 + 5^3 + \ldots\ldots\ldots\ldots\ldots$	$= N^2(2N^2-1)$
(13)	$2 + 4 + 6 + \ldots\ldots\ldots\ldots\ldots\ldots$	$= N(N+1)$
(14)	$2^2 + 4^2 + 6^2 + \ldots\ldots\ldots\ldots\ldots$	$= 2N(N+1)(2N+1)/3$
(15)	$2^3 + 4^3 + 6^3 + \ldots\ldots\ldots\ldots\ldots$	$= 2N^2(N+1)^2$
(16)	$\cdots + 2 + 1 + 0 + 1 + 2 + \ldots\ldots\ldots\ldots$	$= (N^2-1)/4$
(17)	$\cdots + 2^2 + 1^2 + 0^2 + 1^2 + 2^2 \ldots\ldots$	$= N(N^2-1)/12$
(18)	$\cdots + 2^3 + 1^3 + 0^3 + 1^3 + 2^3 + \ldots\ldots\ldots$	$= (N^2-1)^2/32$
(19)	$\cdots + 1.5 + 0.5 + 0.5 + 1.5 + \ldots\ldots\ldots$	$= N^2/4$
(20)	$\cdots + 1.5^2 + 0.5^2 + 0.5^2 + 1.5^2 + \ldots\ldots$	$= N(N^2-1)/12$
(21)	$\cdots + 1.5^3 + 0.5^3 + 0.5^3 + 1.5^3 + \ldots\ldots$	$= N^2(N^2-2)/32$
(22)	$\cdots + 3 + 1 + 1 + 3 + \ldots\ldots\ldots\ldots$	$= N^2/2$
(23)	$\cdots + 3^2 + 1^2 + 1^2 + 3^2 + \ldots\ldots\ldots$	$= N(N^2-1)/3$
(24)	$\cdots + 3^3 + 1^3 + 1^3 + 3^3 + \ldots\ldots\ldots$	$= N^2(N^2-2)/4$
(25)	$\cdots + 4 + 2 + 0 + 2 + 4 \ldots\ldots\ldots$	$= (N^2-1)/2$
(26)	$\cdots + 4^2 + 2^2 + 0^2 + 2^2 + 4^2 \ldots\ldots\ldots$	$= N(N^2-1)/3$
(27)	$\cdots + 4^3 + 2^3 + 0^3 + 2^3 + 4^3 \ldots\ldots\ldots$	$= (N^2-1)^2/4$
(28)	$1 + 3 + 5 + 3 + 1 \ldots\ldots\ldots\ldots\ldots$	$= (N^2+1)/2$
(29)	$1^2 + 3^2 + 5^2 + 3^2 + 1^2 \ldots\ldots\ldots\ldots$	$= N(N^2+2)/3$
(30)	$1^3 + 3^3 + 5^3 + 3^3 + 1^3 \ldots\ldots\ldots\ldots$	$= (N^4+4N^2-1)/4$
(31)	$1 + 3 + 5 + 5 + 3 + 1 \ldots\ldots\ldots\ldots$	$= N^2/2$
(32)	$1^2 + 3^2 + 5^2 + 5^2 + 3^2 + 1^2 \ldots\ldots\ldots$	$= N(N^2-1)/3$
(33)	$1^3 + 3^3 + 5^3 + 5^3 + 3^3 + 1^3 \ldots\ldots\ldots$	$= N^2(N^2-2)/4$

Moments.—The term *moment* is applied to the means of powers of deviations of a given set of numbers about a specified origin. If the deviations (d) are taken from an arbitrary origin they are designated as follows:

First moment, $v_1 = \Sigma d \div N$

Second moment, $v_2 = \Sigma d^2 \div N$

Third moment, $v_3 = \Sigma d^3 \div N$

Fourth moment, $v_4 = \Sigma d^4 \div N$
etc.

If the moments are taken about the mean as origin, the symbol m may be substituted for v. Obviously $m_1 = 0$, and m_2 is the variance. In a normal distribution, or in a series having a constant first difference, $m_3 = 0$. The Greek letter μ may be used to indicate moments about the mean instead of m. However, it is also used to indicate moments of a distribution, expressed in units of class intervals, and adjusted for grouping errors by Sheppard's correction, in which case, when $i = 1$,

$$\mu_2 = m_2 - \frac{1}{12}$$

$$\mu_4 = m_4 - \frac{m_2}{2} + \frac{7}{240}$$

A derived measure of skewness is

$$\beta_1 = \mu_3^2/\mu_2^3$$

NOTES ON CHAPTERS XIV-XVI

SIMPLE CORRELATION

Proof of correlation variances.—The equality of $\sigma_y^2 = \sigma_d^2 + \sigma_t^2$ or $\Sigma y^2 = \Sigma t^2 + \Sigma d^2$ may be verified as follows:

Defining, expanding, simplifying, and assuming $\Sigma Y = \Sigma T$ and $\Sigma YT = \Sigma T^2$:

(1)
$$\Sigma y^2 = \Sigma \left(Y - \frac{\Sigma Y}{N} \right)^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}$$

(2)
$$\Sigma t^2 = \Sigma \left(T - \frac{\Sigma Y}{N} \right)^2 = \Sigma T^2 - \frac{(\Sigma Y)^2}{N}$$

(3)
$$\Sigma d^2 = \Sigma (Y - T)^2 = \Sigma Y^2 - \Sigma T^2$$

$\Sigma Y = \Sigma T$ by first normal equation.
$\Sigma YT = \Sigma T^2$ is demonstrated below.
Hence, adding (2) and (3), $\Sigma y^2 = \Sigma t^2 + \Sigma d^2$.
To show $\Sigma YT = \Sigma T^2$ for $T = a + bX$, square T by guide factors:

	a	bX	Summing
a	$a^2 + abX$		$na^2 + ab\Sigma X$
bX	$abX + b^2X^2$		$ab\Sigma X + b^2\Sigma X^2$

with $= T^2$ and $= \Sigma T^2$.

By comparison with the normal equations it will be seen that the summed first line totals $a\Sigma Y$ and second line $b\Sigma XY$, which is identical with $\Sigma YT = \Sigma Y(a + bX) = a\Sigma Y + b\Sigma XY$. The same proof may be generalized by the relation of T^2 to the normal equations for the parabola, cubic, etc.

It may be deduced from the above that, for the parabola:

$$\Sigma t^2 = \Sigma T^2 - \frac{(\Sigma Y)^2}{n} = a\Sigma Y + b\Sigma XY + c\Sigma X^2 Y - \frac{(\Sigma Y)^2}{N}$$

$$\Sigma d^2 = \Sigma Y^2 - \Sigma T^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY - c\Sigma X^2 Y$$

and σ_t and σ_y may be obtained from these expressions. For the straight-line trend, terms in c may be dropped, and for the cubic, etc., the series is extended to terms in d, etc. The same proof may be written in the more general form

$$T = a + b_1 X_1 + b_2 X_2$$

in which case

$$\Sigma t^2 = b_1 \Sigma x_1 y + b_2 \Sigma x_2 y$$

The form may be applied to any number of elements. These equations are often written with X_0 and x_0 replacing Y and y, respectively.

Other features of correlation.—There are a number of other, less important propositions in connection with correlation which, for the most part, are fairly obvious and scarcely require formal proof. The first of these propositions is that r^2 is limited to the range 0 to 1, which means that r may range from $+1$ to -1. This obviously follows from the proposition just stated that

$$\sigma_t^2 + \sigma_d^2 = \sigma_y^2$$

which, divided by σ_y^2, gives

$$r^2 + k^2 = 1$$

Since both r^2 and k^2 are necessarily positive, if they have any value above zero, the proposition just stated is obvious.

There are several convenient equalities which also are obvious and merely require stating. For example,

$$\Sigma xy = \Sigma x(Y - M_Y) = \Sigma xY - M_Y\Sigma x$$

These equalities are obtained by expanding the terms $(Y - M_y) \times x$ and attaching the Σ sign to the variables. But since $\Sigma x = 0$, the equalities reduce simply to the statement

$$\Sigma xy = \Sigma xY$$

Another series of equalities begins with

$$r^2 = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} = b\frac{\Sigma xy}{\Sigma y^2} = b^2\frac{\Sigma x^2}{\Sigma y^2} = b^2\frac{\sigma_x^2}{\sigma_y^2}$$

These equalities are readily obtained as follows: The first fraction is merely the formula for r^2; the second extracts from the fraction $b = \Sigma xy/\Sigma x^2$; the third squares b and divides the fraction by $\Sigma xy/\Sigma x^2$; and the fourth reduces the fraction by dividing both terms by N. The square root of the last expression is

$$r = b\frac{\sigma_x}{\sigma_y}$$

<div align="center">

EXAMPLE A5

INTERCORRELATIONS AND REGRESSIONS [1]

</div>

Data: Assumed correlated series, W, X, and Y.

		W	X	Y	Z	
		1	8	2	11	
		2	5	4	11	
		2	4	10	16	
		5	1	20	26	
	Σ	10	18	36	64	
P	W	34	31	130	195	
	X		106	96	233	
	Y			520	746	
	Z				1,174	
Np	W	36	-56	160	140	
	X	-56	100	-264	-220	
	Y	160	-264	784	680	
	Z				600	$(-M)$
b	W	1.00000	-1.55556	4.44444	3.88889	-2.5
	X	-0.56000	1.00000	-2.64000	-2.20000	-4.5
	Y	0.20408	-0.33673	1.00000	0.86735	-9.0
M		2.5	4.5	9.0	16.0	
a	W	0	8.38889	-2.11111	6.27778	
	X	5.02000	0	20.88000	25.90000	
	Y	0.66328	7.53057	0	8.19385	
r^2	W	1.00000	0.8711	0.9070		
	X		1.00000	0.8890		
	Y			1.0000		

[1] *Explanation*: The two series from which each computed item below Σ is derived are indicated by column and row designations, e.g., an item in column X, row W, is derived from X and W series. Block P contains summed cross products, e.g., $34 = \Sigma W^2$; $31 = \Sigma W X$. Np contains N times centered cross products, e.g., $36 = N\Sigma w^2 = N\Sigma W^2 - (\Sigma W)^2$; $-56 = N\Sigma wx = N\Sigma WX - \Sigma W\Sigma X$. The b's are the Np's, each divided by the summed squares of its row; that is, the diagonal numbers 36, 100, and 784 are the divisors. Each a is the M in the same column combined with b (of similar position as a) times negative M at right. Each a and b pair (related by similar positions) describes the regression of the series at head of column on the series indicated by the row designation at left (e.g., regression of W on X is $5.02 - 0.56 X$; and the regression of Y on W is $-2.11 + 4.44W$. Each squared r is the product of two related b's; e.g., squared r of series W and X is $(-1.55)(-0.56) = 0.8711$. Z checks apply in all blocks except the last one. Blocks containing r's and F's may readily be added.

If the terms are then rearranged, b is available as

$$b = r\frac{\sigma_y}{\sigma_x} \quad \text{and} \quad b_1 = \beta_1\left(\frac{\sigma_y}{\sigma_{x_1}}\right)$$

Correlation by diagonal deviations.—The linear coefficient of correlation (r) as computed from tabulated data may be found by means of a frequency distribution based upon the diagonals of the scatter tabulation, assuming that the X and Y scales are expressed in unit intervals. The computation is based upon the equation

$$2(\sigma_x^2 + \sigma_y^2) = (\sigma_\theta^2 + \sigma_w^2)$$

where σ_θ^2 is the variance of the distribution having frequencies obtained by adding the diagonals from the upper left to the lower right taken as columns, and σ_w^2 is the variance of the distribution having frequencies obtained by adding the diagonals from the upper right to the lower left, taken as columns. σ_x and σ_y are the standard deviations of the columns and rows, respectively, as in other correlation problems. Class intervals in each case are taken as unity. The factor 2 in the equation arises from the fact that the diagonal scales are really $\sqrt{2}$ times the x and y scales. It may also be shown that $\sigma_\theta^2 - \sigma_w^2 = 4\Sigma xy/N$.

It follows from this equality that r may be found by the formula

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_w^2}{2\sigma_x\sigma_y} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{X-Y}^2}{2\sigma_x\sigma_y}$$

If, however, σ_θ^2 is substituted for σ_w^2, the r thus obtained will be the same in magnitude, but the sign will be reversed. Hence, it is convenient to use σ_θ^2 with negative correlations and σ_w^2 with positive correlations, since these diagonals are more quickly obtainable.

Intercorrelations: The r coefficients relating each of several series to the other series, respectively, is often required. A convenient form for computing and tabulating such correlations, together with the regressions, appears in Example A5. Brief explanations appear in a footnote below. In reading results, the dependent series is read in the column headings and the independent in the row designations. For example, the slope (block b) of X on W is in column X, row W ($b = -1.55$). The blocks P and N_p are as explained in multiple correlation, below.

Doolittle solution of multiple correlation.—In Example A6 a problem in multiple correlation with three independent series is solved by a method adapted to laboratory practice. The data are summed by rows (column Z), and the columns, including Z, are summed in row S. Cross products are then tabulated in block P: the first row lists ΣX_1^2, ΣX_1X_2, ΣX_1X_3, ΣX_1X_0; the second row lists ΣX_2^2, ΣX_2X_3, etc. The subscripts in column and row designations indicate each cross product. These cross products may be run off on a calculator, two or three at a time, where the multipliers are common.

In block Np the centered cross products are similarly listed, each multiplied by N to avoid decimals. The centering utilizes formulas such as

$$N\Sigma x_1^2 = N\Sigma X_1^2 - \Sigma X_1\Sigma X_1$$

$$N\Sigma x_1x_2 = N\Sigma X_1X_2 - \Sigma X_1\Sigma X_2, \text{ etc.}$$

each of which may be solved in one operation on a calculator.

In both P and Np blocks, the Z column furnishes a convenient check. In each of these blocks Z is entered as the sum of the "full rows" including items in block column above first listed item of each row. These are the items appropriate to the unfilled spaces in the lower left-hand portion of the block. The block footings of the Z items thus found should then check as ΣZ^2 in the P block, and as $N\Sigma z^2$ in the Np block. In locating errors, each Z may be found as if the data row sums under Z were another X series.

The Doolittle solution begins in block I. Row p_1 is entered and is pointed off for convenience to four decimals, as are other items in the Np block, when entered below. The row thus entered is labeled s_1. It is divided by the first item (0.2400), and signs changed, to obtain row q_1.

Block II similarly begins with row $p_2 \div 1,000$. The next row is s_1 (beginning in column 2) multiplied by the q_1 item of column 2. Rows 1 and 2 are then added to obtain row s_2 and q_2 is obtained by dividing by -0.7396. In these calculations the Z column is included, and the Z item of row s_2 should check as the row sum, except for small rounding errors.

Block III is similarly obtained. Row 1 is row $p_3 \div 1,000$. Row 2 is row s_1, beginning in the third column, times the q_1 item of the same column. Row 3 is similarly obtained from row s_2 times q_2 of column 3. The three rows are then summed to obtain s_3, which, as before, is divided by its first item, and signs changed, to obtain q_3.

In block C the constants of the regression equation are derived from the q rows. These rows are brought down for convenience, though this is not strictly necessary. The b's are found in reverse order. In column 0, the item -1 is entered and may be regarded as b_0. Each q of column 0 is then multiplied by -1, and the results are listed below. The last item thus found (0.4360) is b_3 and is so entered in the b row. This b is then multiplied by the q's above it, and the results entered below in appropriate rows, as indicated. The sum of bq_2 row is b_2, and is so entered in the b row. It in turn is then multiplied by the q above it, completing row bq_1, whose sum is b_1. Thus the rows below b may be described as bq_1, bq_2, and bq_3, though the computation is piecemeal, from right to left. When more series are involved, the blocks of the solution, of course, are expanded, but the method remains the same.

The constant a may be discovered by the formula (derived from first normal equation),

$$Na = \Sigma Y - b_1\Sigma X_1 - b_2\Sigma X_2 - b_3\Sigma X_3$$

or the same computation may, in effect, be carried out by multiplying the data sums (row S) by the b's, including -1 taken as b_0. The sum of these products is $-Na$, which divided by $-N$ gives a.

The coefficient R may similarly be found by formula as

$$R^2 = (b_1\Sigma x_1 x_0 + b_2\Sigma x_2 x_0 + b_3\Sigma x_3 x_0) \div \Sigma x_0^2$$

or the same operation may be carried out by multiplying the full row p_0 by row b, as indicated. A useful check is $\Sigma bP_1 = 0$, etc.

SIGNIFICANCE OF THE BETAS

In Example A7 the Doolittle method as applied to multiple correlation is extended to include estimates of the significance of the betas. In this process it is necessary to begin with the zero order coefficients, or r's, instead of the centered cross products.

EXAMPLE A6

MULTIPLE CORRELATION, DOOLITTLE SOLUTION

Data: Assumed records of workers A, B, C, etc., in respect to entrance tests, experience (years), education (grades), and efficiency as measured by sales (thousands of dollars).

	Block	Row	X_1 Test Score	X_2 Experience	X_3 Education	X_0 Efficiency	Z Sums or checks
Data	D	A	4	5	6	5	20
		B	5	2	7	4	18
		C	6	4	5	5	20
		D	4	9	8	6	27
		E	5	8	8	9	30
		F	6	4	11	10	31
		G	6	10	7	9	32
		H	7	11	8	12	38
		I	9	10	8	11	38
		J	8	7	12	9	36
Data sums		S	60	70	80	80	290
Cross products	P	P_1	384	437	493	508	1822
		P_2		576	566	614	2193
		P_3			680	668	2407
		P_0				710	2500
		Z					$\Sigma Z^2 = 8922$
Centered cross products, times N	Np	p_1	240	170	130	280	820
		p_2		860	60	540	1630
		p_3			400	280	870
		p_0				700	1800
		z					$N\Sigma z^2 = 5120$

			0.2400	0.1700	0.1300	0.2800	0.8200
$p_1/1{,}000$	I	s_1	0.2400	0.1700	0.1300	0.2800	0.8200
$-s_1/0.2400$		q_1	-1	-0.7083	-0.5417	-1.1667	-3.4167 Check
$p_2/1{,}000$	II	1		0.8600	0.0600	0.5400	1.6300
$s_1(-0.7083)$		2		-0.1204	-0.0921	-0.1983	-0.5808 Check
Add		s_2		0.7396	-0.0321	0.3417	1.0492
$-s_2/0.7396$		q_2		-1	0.0434	-0.4620	-1.4186 Check
$p_3/1{,}000$	III	1			0.4000	0.2800	0.8700
$s_1(-0.5417)$		2			-0.0704	-0.1517	-0.4442
$s_2(0.0434)$		3			-0.0014	0.0148	0.0455
Add		s_3			0.3282	0.1431	0.4713 Check
$-s_3/0.3282$		q_3			-1	-0.4360	
Copy q_1	C	q_1		-0.7083	-0.5417	-1.1667	
Copy q_2		q_2			0.0434	-0.4620	
Copy q_3		q_3				-0.4360	
Enter -1		b	0.5899	0.4809	0.4360	-1.0000	
Rows $b \times q_1$	R	bq_1		-0.3406	-0.2362	$+1.1667$	$=$ Row $\Sigma = b_1$
Rows $b \times q_2$		bq_2			0.0189	$+0.4620$	$=$ Row $\Sigma = b_2$
Rows $b \times q_3$		bq_3				0.4360	$= b_3$
Rows $b \times S$		$-Na$	35.394	$+33.663$	$+34.880$	-80.000	$= 23.937$
$\div(-N)$		a					$= -2.3937$
Rows $b \times p_1$			141.576	$+81.753$	$+56.680$	-280	$= .009$ Check
Rows $b \times p_2$			100.283	$+413.574$	$+26.160$	-540	$= .017$ Check
Rows $b \times p_3$			76.687	$+28.854$	$+174.400$	-280	$= .059$ Check
Rows $b \times p_0$	R	$-N\Sigma d^2$	165.172	$+259.686$	$+122.080$	-700	$= -153.062$
$-N\Sigma d^2/(-700)$		k^2					$= 0.2187$
$1 - k^2$		R^2					$= 0.7813$
		F	$(0.7813/0.2187)$	$(6/3)$			$= 7.14$

Example A7

MULTIPLE CORRELATION WITH SIGNIFICANCE OF BETAS

Data: See preceding example.
A. Zero-order coefficients (r's).

I. Centered cross-products

Block	Row	X_1	X_2	X_3	X_0	Z
Np	p_1	240	170	130	280	820
	p_2	170	860	60	540	1630
	p_3	180	60	400	280	870
	p_0	280	540	280	700	1800

II. Linear slopes (X of row number, independent)

Block	Row	X_1	X_2	X_3	X_0	Z
b	1	$b_{11} = 1.0000$	$b_{21} = 0.7083$	$b_{31} = 0.5417$	$t_{01} = 1.1667$	3.4167
	2	$b_{12} = 0.1977$	$b_{22} = 1.0000$	$b_{32} = 0.0698$	$b_{02} = 0.6279$	1.8953
	3	$b_{13} = 0.3250$	$b_{23} = 0.1500$	$b_{33} = 1.0000$	$b_{03} = 0.7000$	2.1750
	0	$b_{10} = 0.4000$	$b_{20} = 0.7714$	$b_{30} = 0.4000$	$t_{00} = 1.0000$	2.5714

III. $r_{12}^2 = b_{12}b_{21}$; $r_{13}^2 = b_{13}b_{31}$, etc.

Block	Row	X_1	X_2	X_3	X_0	Z
r^2	1	1.0000	0.1400	0.1761	0.4667	
	2		1.0000	0.0105	0.4844	
	3			1.0000	0.2800	
	0				1.0000	

B. Solution, including significance of betas.

Operations			Correlation of:				Significance of:			Z
			X_1	X_2	X_3	X_0	β_1	β_2	β_3	
Square roots of Block r^2	r	r_1	1.	0.3742	0.4196	0.6831	1.			3.4769
		r_2		1.	0.1023	0.6960		1.		3.1725
		r_3			1.	0.5292			1.	3.0511
		r_0				1.				
Enter r_1	I	s_1	1.	0.3742	0.4196	0.6831	1.			3.4769
$-s_1$		g_1	-1.	-0.3742	-0.4196	-0.6831	-1.			-3.4769
Enter r_2	II	1		1.	0.1023	0.6960		1.		3.1725
$s_1 \times$ second g_1		2		-0.1400	-0.1570	-0.2556	-0.3742			-1.3011
Sum		s_2		0.8600	-0.0547	0.4404	-0.3742	1.		1.8714
$-s_2 \div$ first s_2		g_2		-1.	0.0636	0.5121	0.4351	-1.1628		-2.1760
Enter r_3	III	1			1.	0.5292			1.	3.0511
$s_1 \times$ third g_1		2			-0.1761	-0.2866	-0.4196			-1.4589
$s_3 \times$ second g_2		3			-0.0035	0.0280	-0.0238	0.0636		0.1190
Sum		s_3			0.8204	0.2706	-0.4434	0.0636		1.7112
$-s_3 \div$ first s_3		g_3			-1.	-0.3298	0.5405	-0.0775	-1.2189	-2.0858
Enter -1, col. 0	C	β	$\beta_1=0.3452$	$\beta_2=0.5331$	$\beta_3=0.3298$	-1.	1.4025	1.1677	1.2189	(1) $= \Sigma|q_i|$
$\beta \times q_1$		β_1		-0.1995	-0.1384	0.6831	0.0511	0.0426	0.0444	(2) $=$ (1) $\times$ R3
$\beta \times q_2$		β_2			0.0210	0.5121	0.2261	0.2064	0.2107	(3) $= \sqrt{(2)}$
$\beta \times q_3$		β_3				0.3298	1.527	2.583	1.565	(4) $= C\beta \div$ (3)
Enter full r_0	R	r_0	0.6831	0.6960	0.5292	1.				
$-\beta \times r_0$		1	-0.2358	-0.3710	-0.1745	$=0.2187$				
Σ last row		2			$1-R^2$	0.03645				
$(1-R^2) \div (N-m)$		3								

The r's are obtained from the centered cross products (times N) which are copied in Block Np from the preceding example. This block is written in complete rows. Each row is divided by the $N\Sigma x^2$ of the row, to obtain the regression slopes indicated, and the r squares are found as the product of the two correlative b's as $b_{21}b_{12}$, etc. The r's are then entered for solution (more decimals were carried in r^2 than are indicated).

To obtain σ_β it is necessary to affix beta columns with an entry of 1 in each column corresponding to the initial 1's of the r's. These entries are included in the Z column
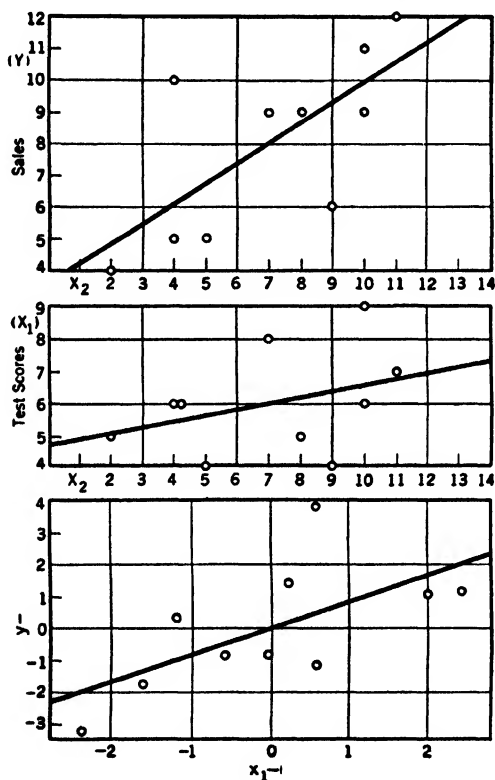
FIG. A1—The Net or Partial Regression of Y on X_1. The regression of Y and X_1 on X_2 are eliminated in obtaining the residual deviations, y- and x_1-, as plotted. (Data: See Example A8.)

at the extreme right. The steps in the solution are as before, except that the β columns are included in each step, parallel with the X columns. In solving the constants the q's are not here brought down, but the solution is as before. The expression $(1 - R^2) \div (N - m)$ is solved for use in discovering the σ_β's.

The beta columns are summed as the absolute products q_1s_1; q_2s_2; etc. These sums in reality are the reciprocals of $1 - R^2_{1.23}$, $1 - R^2_{2.13}$, and $1 - R^2_{3.12}$, and the

<div align="center">

EXAMPLE A8

THEORETICAL ANALYSIS OF PARTIAL CORRELATION

</div>

Data: Sales records in units per week (Y), psychological-test scores (X_1), and years of experience (X_2) of a group of salesmen. Subscripts 0, 1, and 2 refer to Y, X_1, and X_2 series, respectively (see Example 16.1).

Sales-men	Sales Y	Test X_1	Years X_2	Regressions		Errors		Product of errors $d_{02} \times d_{12}$
				Y on X_2 T_{02}	X_1 on X_2 T_{12}	d_{02} $Y - T_{02} =$	d_{12} $X_1 - T_{12}$	
A	5	4	5	6.7442	5.6047	−1.7442	−1.6047	2.7989
B	4	5	2	4.8605	5.0116	−0.8605	−0.0116	0.0099
C	5	6	4	6.1163	5.4070	−1.1163	0.5930	−0.6619
D	6	4	9	9.2558	6.3953	−3.2558	−2.3953	7.7986
E	9	5	8	8.6279	6.1977	0.3721	−1.1977	−0.4456
F	10	6	4	6.1163	5.4070	3.8837	0.5930	2.3030
G	9	6	10	9.8837	6.5930	−0.8837	−0.5930	0.5240
H	12	7	11	10.5116	6.7907	1.4884	0.2093	0.3115
I	11	9	10	9.8837	6.5930	1.1163	2.4070	2.6869
J	9	8	7	8.0000	6.0000	1.0000	2.0000	2.0000
Σ items	80	60	70	80.0000	60.0000	0.0000	0.0000	17.3254
Σ squares	710	384	576			36.0930	20.6395	

$$b_{01.2} = \frac{\Sigma(d_{02} \times d_{12})}{\Sigma d_{12}^2} \qquad = \frac{17.3254}{20.6395} \qquad = 0.8394$$

$$\beta_{01.2} = b_{01.2}\left(\frac{\sigma_x}{\sigma_y}\right) \qquad = (0.8394)\left(\frac{1.5492}{2.6458}\right) \qquad = 0.4915$$

$$r_{01.2} = \frac{\Sigma(d_{02} \times d_{12})}{(\Sigma d_{02}^2 \Sigma d_{12}^2)^{1/2}} \qquad = \frac{17.3254}{(36.0930)(20.6395)^{1/2}} \qquad = 0.635$$

$$r_{01.2} = \frac{r_{01} - r_{02}r_{12}}{[(1 - r_{02}^2)(1 - r_{12}^2)]^{1/2}} \qquad = \frac{0.683 - 0.696 \times 0.374}{[(1 - 0.696^2)(1 - 0.374^2)]^{1/2}} = 0.635$$

$$t_{01.2} = \frac{\Sigma(x_1 - b_{12}x_2)(x_0 - b_{02}x_2)}{\Sigma(x_1 - b_{12}x_2)^2}\ (x_1 - b_{12}x_2)^2$$

$$F = \frac{r_{01.2}^2}{1 - r_{01.2}^2} \times \frac{N - m}{m - c} = \frac{0.403}{0.597} \times \frac{10 - 3}{3 - 2} = 4.73$$

where c is the number of coefficients (b_{12} and b_{02}) used *indirectly* (i.e., other than the usual coefficients of the regression) to determine the regression. Tabular $F = 5.59$ and 12.25 for $DF = 1$ and 7.

method of their solution, thus illustrated, is much more economical than direct solution. Multiplying each by $(1 - R^2) \div (N - m)$ gives each σ_β^2, as may be seen by reference to its formula (cf. page 384), and the square root is σ_β. The ratio of each β to its σ_β is t, which may be evaluated by reference to the table of t, p. 586.

In using the method of multiple correlation, beginning with the r's as just described, it will be seen that the regression equation is not directly obtained. It can, however, be computed by the formulas

$$b_1 = \beta_1 \times \frac{\sigma_0}{\sigma_1} = \beta_1 \sqrt{\frac{\Sigma x_0^2}{\Sigma x_1^2}}$$

$$b_2 = \beta_2 \times \frac{\sigma_0}{\sigma_2} = \beta_2 \sqrt{\frac{\Sigma x_0^2}{\Sigma x_2^2}}, \quad \text{etc.}$$

After the b's have thus been obtained, a may be calculated as in preceding problems.

Analysis of partial correlation.—Partial correlation, worked out in detail for purposes of elucidation, rather than by a general formula, is presented in Example A8 (see Fig. A1). In this example, the partial correlation $r_{01.2}$ is worked out. The regressions of Y (or X_0) on X_2 and X_1 on X_2, and in each case the errors of estimate, are noted. The slope of the former on the latter (d_{02} on d_{12}) is the net regression coefficient, which, if multiplied by σ_x/σ_y, is the corresponding β. The correlation of the two sets of errors of estimate thus obtained is the coefficient of partial correlation. That is, theoretically it is the correlation of sales with the psychological test scores after each of these two sets of data has been corrected for the effects of experience. The coefficient of partial correlation thus obtained does not measure up to the required standard of reliability and may be due merely to chance variability. The measurement of partial correlation, however, is more conveniently carried out by means of equations given in Chapter XVI.

NOTES ON CHAPTER XVII

CURVILINEAR CORRELATION

General rule for decoding data.—From the discussion in Chapter XVII it will be seen that the general rule for decoding a trend equation, expressed in the X and Y scales employed in computing, so as to adapt it to $\overline{X}$ and $\overline{Y}$ scales representing the original data or any new required scales, may be stated as follows:

1. Take R_x and R_y to represent the values on the $\overline{X}$ and $\overline{Y}$ scales, respectively, at the point of origin of the scales X and Y used in calculation.

2. Take i_x and i_y to represent the number of units in the $\overline{X}$ and $\overline{Y}$ scales corresponding to a unit in the X and Y scales, respectively, used in calculation. Obviously in some cases these values might be fractional.

3. Designate the adjusted constants applicable to scale $\overline{X}$ and $\overline{Y}$ by the symbols $\bar{a}$, $\bar{b}$, $\bar{c}$, etc., the symbols a, b, c, etc., being employed in the equation as originally computed.

The rules for decoding in any type of trend equation may then be stated as follows:

1. Multiply the expression representing T, as calculated, by i_y, and then add R_y.

2. For each X appearing in the equation substitute $\dfrac{\overline{X} - R_x}{i_x}$.

3. Simplify as far as possible the complex algebraic expression thus obtained.

The foregoing procedure may generally be reduced to comparatively simple rules for any specific trend equation. While it is not feasible to do this for all trend equations, the rules for a straight line and parabola may be presented.

In conformity with the foregoing notation and procedure, a straight-line trend equation calculated in terms of X and Y may be transformed so as to be applicable to the scales $\overline{X}$ and $\overline{Y}$ by the following formulas:

$$\overline{a} = R_y + i_y \left(a - R_x \frac{b}{i_x} \right)$$

$$\overline{b} = i_y \frac{b}{i_x}$$

Or the procedure may be reduced to the following two steps:

1. Write a and b/i_x, and multiply each by i_y.
2. Employing a and b thus partially corrected,

$$\overline{a} = R_y + a - R_x b$$

$$\overline{b} = b$$

Similarly, an equation for a parabola trend may be decoded or transformed in accordance with the preceding notation, as follows:

$$\overline{a} = R_y + i_y \left(a - R_x \frac{b}{i_x} + R_x^2 \frac{c}{i_x^2} \right)$$

$$\overline{b} = i_y \left(\frac{b}{i_x} - 2R_x \frac{c}{i_x^2} \right)$$

$$\overline{c} = i_y \frac{c}{i_x^2}$$

Or the procedure may be reduced to the following two steps,

1. Write a, b/i_x, and c/i_x^2, and multiply each by i_y.
2. Employing a, b, and c thus partially corrected,

$$\overline{a} = R_y + a - R_x b + R_x^2 c$$

$$\overline{b} = b - 2R_x c$$

$$\overline{c} = c$$

When decoding is very complex, the following procedure may be substituted, or charts may be employed. Code the items on which the estimate is based and obtain the corresponding T. This estimate in terms of T may be simply decoded as $i_y T + R_y$.

Biserial r and biserial η.—As has been indicated in the text, it is frequently desired to measure covariation between variables of a type in which one or both are

expressed in non-quantitative terms, generally to determine whether or not such covariation as may exist is statistically significant. Where both variables are so expressed, this result may be achieved by means of the coefficient of mean square contingency, described on page 427, and the measurement thus obtained may be evaluated by reference to tables or charts of chi-square values.

Where one variable is quantitatively measured and the other appears as a dichotomy, i.e., having but two classes, the procedures of biserial r or biserial η may be used for this purpose. Both devices assume normality in the non-quantitative variable and, upon this assumption, make use of the known characteristics of the normal distribution to measure covariation. In each case, this measurement is effected through reference to tables of A and z, such as those on pages 582 to 585, in which A refers to areas under various portions of the normal curve and z represents a measure of the ordinates at specified points along the base line.

The method of biserial r is illustrated in Example A9, where data are arranged to permit measurement of a possible relationship between the number of workers participating in producing units of a certain commodity and their acceptance or rejection by the inspection department of a manufacturing concern. Acceptances and rejections are designated as f_2 and f_1, respectively, being classified according to the number of operatives involved in processing. The totals of these frequencies are designated as f. Ordinary procedure first calculates the mean of the smaller of the two classes (the *tail* of the distribution) and then notes the mean and the standard deviation of the whole distribution (columns $f_1 Y$, fY, and fY^2). It notes, also, the proportion of the whole non-quantitative distribution in the *tail* and designates this proportion as q (in contrast with which the larger portion may be regarded as p). The appropriate z value is then taken from tables of q and z, or if the latter are not available, from tables of A and z, in which case the proper value of A is determined by the fact that $A = (1 - 2q)/2$. This relationship may be explained by the fact that the area referred to is that under one-half the normal curve less that represented by the proportion of the whole distribution that features its tail. The value of r_{bis} is then available directly by formula, as indicated in the example.

The method of biserial η differs principally in that it avoids reference directly to the standard deviation of the quantitatively measured series. As indicated in Example A10, in which the same data are used, the two classes of the dichotomy are described as f_1 and f_2 as before, and the totals of these classes are noted as f. In this case, however, reference is made to the portion of the whole that is represented by the larger of the two classes, f_2, and the ratio of these frequencies to total frequencies is designated as T. T values are reduced to those representing one-half a normal distribution (since only the latter are usually described in tables of areas and ordinates) by subtracting 0.50 from each T value. Distances from the mean ordinate, designated as x or x/σ, for each A are then read from the table, and these values are squared and multiplied by the frequencies for each row. The measurement of biserial η is accomplished by comparison of the x^2 value for the ratio of total frequencies in the major class to those in the whole (designated as B^2) with the total of row products, $\Sigma f(x^2)$. Ordinarily, this total, divided by N, is described as A^2, and that terminology is followed in the example.

The coefficients thus determined may be evaluated as to reliability and significance by reference to their standard errors, or they may be compared with the charts, pages 559 and 560. In using the charts, the number of variables should be regarded as 2 for biserial r and m for biserial η.

CORRELATION BY MEANS OF BISERIAL r

Data: Y, the number of operatives having a part in the production of certain units; X, disposition of these units in the inspection division.

Operatives involved in production (coded) Y	Disposition by inspectors			f_1Y	(ΣY) fY	(ΣY^2) fY^2
	Rejected f_1	Accepted f_2	Total f			
0	70	110	180	0	0	0
1	65	130	195	65	195	195
2	50	210	260	100	520	1,040
3	35	450	485	105	1,455	4,365
4	10	660	670	40	2,680	10,720
5	10	840	850	50	4,250	21,250
Totals:	240	2,400	2,640	360	9,100	37,570
Means:				$M_1 = 1.5$	$M = 3.447$	
Proportions:	0.0909 (q)	0.9091 (p)				

For σ_y:

$$\Sigma y^2 = \frac{N\Sigma Y^2 - \overline{\Sigma Y^2}}{N} = \frac{(2,640 \times 37,570) - (9,100)^2}{2,640} = 6,202.58$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{6,202.58}{2,640}} = 1.53.$$

Biserial r:

A (for reference in table of area and ordinates)

$$= \frac{1 - 2q}{2} = \frac{1 - 2 \times 0.0909}{2} = 0.4091;$$

z (from table, when $A = 0.4091$) $= 0.1636$.

$$r_{bis} = \frac{(M_1 - M)q}{z\sigma_y} = \frac{(1.50 - 3.45)(0.0909)}{(0.1636)(1.53)} = -0.71$$

Standard error:

$$\sigma_{r\ bis} = \frac{\dfrac{\sqrt{pq}}{z} - r^2}{\sqrt{N}} = \frac{\dfrac{\sqrt{(0.0909)(0.9191)}}{0.1636} - (-0.708)^2}{\sqrt{2,640}} = 0.006$$

<div align="center">

EXAMPLE A10

CORRELATION BY MEANS OF BISERIAL ETA

</div>

Data: Same as Example A9.

Operatives involved in production (coded) Y	Disposition by inspectors			$T = \dfrac{f_2}{f}$	$A = T - 0.50$	x	x^2	fx^2
	Re-jected f_1	Ac-cepted f_2	Total f					
0	70	110	180	0.6111	0.1111	0.2823	0.0797	14.346
1	65	130	195	0.6667	0.1667	0.4307	0.1856	36.173
2	50	210	260	0.8077	0.3077	0.8696	0.7560	196.560
3	35	450	485	0.9278	0.4278	1.4596	2.1304	1,033.244
4	10	660	670	0.9851	0.4851	2.1727	4.7206	3,162.802
5	10	840	850	0.9882	0.4882	2.2635	5.1234	4,354.890
	240	2,400	2,640	0.9091	0.4091	1.3346	1.7812 $= B^2$	8,796.995

$$A^2 = \frac{\Sigma(fx^2)}{N} = \frac{8796.995}{2640} = 3.3322$$

$$B^2 \text{ (from } x^2 \text{ column)} = 1.7812$$

$$\eta^2 = \frac{A^2 - B^2}{1 + A^2} = \frac{1.5510}{4.3322} = 0.358$$

$$\eta = \sqrt{.358} = 0.598 \text{ (negative by inspection)}$$

Intraclass correlation.—Problems in correlation sometimes appear in which it is impossible to determine which of two paired items should be placed in the X and which in the Y column. For example, suppose that test scores are available measuring a specific ability of pairs of brothers, the age or other distinguishing features not being recorded. The question arises whether ability of brothers is correlated.

In such cases, one possible method of solution enters the scores of each pair of brothers twice, alternately in the X and Y columns. To illustrate in Example A11, the first pair of brothers made scores of 10 and 9, respectively. Each score may be entered as $X = 10$ and $Y = 9$, also as $X = 9$ and $Y = 10$. Other paired items may be similarly treated. The correlation process is then continued in the usual way, with ten items in each column. Or, with the five pairs listed arbitrarily as X and Y,

$$r = \frac{2\Sigma XY - (\Sigma X + \Sigma Y)^2/2N}{\Sigma X^2 + \Sigma Y^2 - (\Sigma X + \Sigma Y)^2/2N}$$

Example A11

INTRACLASS CORRELATION

Data: Assumed test scores of pairs of brothers, families A, B, C, etc. Correlation tested by single variance.

Family	A	B	C	D	E	Row totals
Scores of	9	11	8	14	16	
brothers (Y)	10	12	10	15	20	
ΣY	19	23	18	29	36	125; $125^2/10 = 1562.5$
ΣY^2	181	265	164	421	656	$1687 - 1562.5 = 124.5$
$(\Sigma Y)^2/N$	180.5	264.5	162	420.5	648	$1675.5 - 1562.5 = 113.0$

MEAN SQUARES

Squares	DF	MS
$\Sigma y^2 = 124.5$	9	
$\Sigma i_c^2 = 113.0$	4	28.25
$\Sigma d^2 = 11.5$	5	2.30

$$F = \frac{28.25}{2.30} = 12.28 \ (1\%F = 11.39)$$

$$\text{Intraclass } r = \frac{2\Sigma i_c^2 - \Sigma y^2}{\Sigma y^2}$$

$$= \frac{2 \times 113 - 124.5}{124.5} = 0.815$$

From the standpoint of the significance of such correlation, however, the problem may best be approached by means of simple variance analysis. It is this method which is illustrated in Example A11. The procedure does not require comment since it parallels that of Example 19·2, page 464. A computation of r by means of a formula is appended.

Graphic multiple curvilinear correlation.—Assuming the following series, u, v, and y, expressed in deviations from the mean, where u and v are taken as independent series and y as a dependent series,

u	v	y
2	0	8
1	-2	-2
-2	2	-10
0	1	5
-1	-1	-1

a linear multiple correlation may be calculated or merely estimated,[1] giving $b_1 = 4$ and $b_2 = 1$, constants for u and v, respectively. The functions thus indicated ($4u$ and v) give the t indicated below. Then find s_1 as $y - t_1$ and $f(u)_1 + s_1$ and $f(v)_1 + s_1$. The results are as follows:

$f(u)_1 + f(v)_1 = t_1$	$y - t_1$ s_1		s_1	$f(u)_1 + s_1$	$f(v)_1 + s_1$	$f(u)_2$	$f(v)_2$
8	0	8	0	8	0	7	3
4	-2	2	-4	0	-6	5	-5
-8	2	-6	-4	-12	-2	-11	-1
0	1	1	4	4	5	2	2
-4	-1	-5	4	0	3	-3	1

The expressions $f(u)_1 + s_1$ and $f(v)_1 + s_1$ are employed in finding revised functions of u and v, respectively (see Fig. A2). This is done by plotting the series against their
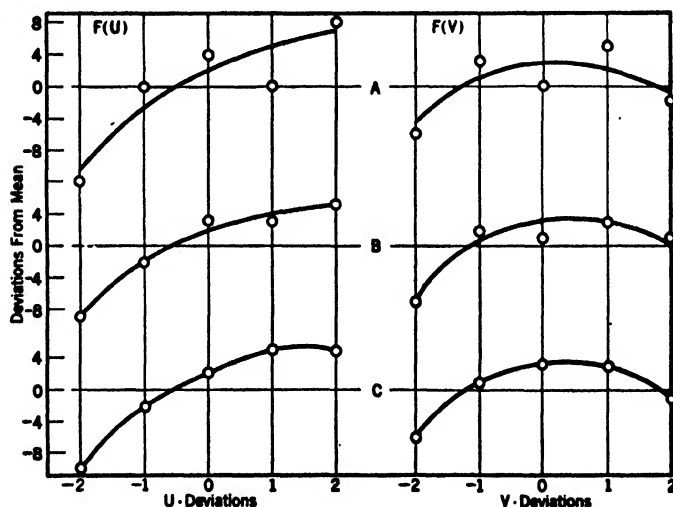
Fig. A2.—Revising Curvilinear Functions. Small circles, first row, are $f(u)_1 + s_1$ and $f(v)_1 + s_1$, respectively, and the smoothed lines give $f(u)_2$ and $f(v)_2$. The second and third rows represent similar third and fourth revisions.

respective deviations, that is, $f(u)_1 + s_1$ against u and $f(v)_1 + s_1$ against v. The functions thus plotted are smoothed and re-estimated by reading items from the charts so that, in each case, the sum is zero. This second estimate of the two functions, as read from the chart, has here been calculated as follows. The differences $y - t_2 = s_2$

---

[1] See page 428 for suggestions regarding such estimates. Also, the coefficients b_1 and b_2 of a multiple linear regression may be estimated by the graphic process of partial correlation, or sometimes they may be approximated by drawing on similar scales the regression of X_1, X_2, and Y on Y.

are again taken and are added to $f(u)_2$ and $f(v)_2$. As before, $f(u)_2 + s_2$ and $f(v)_2 + s_2$ are plotted against u and v, respectively, to obtain a third estimate of $f(u)$ and $f(v)$.

$f(u)_2 + f(v)_2 = t_2$			$y - t_2$ / s_2	$f(u)_2 + s_2$	$f(v)_2 + s_2$	$f(u)_3$	$f(v)_3$
7	3	10	−2	5	1	5	3
5	−5	0	−2	3	−7	4	−7
−11	−1	−12	2	−9	1	−9	0
2	2	4	1	3	3	2	3
−3	1	−2	1	−2	2	−2	1

Similarly, a fourth estimate of the two functions was obtained. In general, this process of approximation is continued until relatively smoothed functions are obtained, if that is possible. While these functions cannot be readily expressed by an equation, an estimate for a new U and V may be read from the final chart and a prediction made accordingly. In so doing the U and V data are first reduced to deviations (u and v) in terms of the means already calculated, and the corresponding t functions are read from the chart and added. To this result M_Y is added to obtain the estimate T comparable to Y.

NOTES ON CHAPTER XVIII

The coefficient of correlation (r) is often used simply as a measure of similarity between two related sets of deviations. However, the Pearsonian r is defined as a measure applicable to normal distributions in the statistical fields sampled. Hence a simpler measure not based upon such an assumption seems desirable. Such a measure, utilizing the average instead of the standard deviation, is available in the so-called coefficient of similarity (symbol Sm).

EXAMPLE A12

COEFFICIENT OF SIMILARITY, Sm

Data: Assumed times series, X and Y, with approximate trends.

Year	X	Y	T_X	T_Y	d_x	d_y	d_x/AD	d_y/AD	s
1923	5	3	4	5	1	−2	0.5	−1.0	−0.5
1924	3	4	6	7	−3	−3	−1.5	−1.5	1.5
1925	7	12	8	9	−1	3	−0.5	1.5	−0.5
1926	13	13	10	11	3	2	1.5	1.0	1.0
1927	10	12	12	13	−2	−1	−1.0	−0.5	0.5
1928	16	16	14	15	2	1	1.0	0.5	0.5
	54	60	54	60	+6	+6	+3.0	+3.0	6)2.5
					−6	−6	−3.0	−3.0	$Sm=0.42$
					6)12	6)12	6)6.0	6)6.0	
					$AD = 2$	$AD = 2$	$AD=1.0$	$AD=1.0$	

The first step in calculating the coefficient of similarity is the reduction of both the X and Y series to units of the average deviation, that is, each X is divided by AD_X

and each Y by AD_Y. The deviations may be from the mean or from the normal, in accordance with the nature of the correlation. As the next step, from each pair of correlated items thus expressed, the numerically smaller (without regard to signs) is selected, and the sign of correlation for that pair is prefixed (like signs give plus; unlike, minus). The algebraic average of the items thus selected (s) is the coefficient of similarity ($Sm = \Sigma s/N$). It should not be assumed that this measure takes no account of the larger deviations, since these are taken account of in N. It may easily be shown that the limits of Sm are the same as the limits of r, that is, ± 1. For a normal correlation surface, the relation of Sm to r is expressed by the formula $r^2 = 2Sm^2 - Sm^4$. The procedure is illustrated by simple assumed data in Example A12 (for short cuts see Davies and Crowder's *Methods of Statistical Analysis*, Chapter VIII).

NOTES ON CHAPTER XX

The reliability of correlation measures.—The convenient appraisal of the significance of correlation has been greatly advanced in recent years by the English statistician R. A. Fisher. Fisher's first step involved the transformation of the r scale, so that its sampling distribution would become approximately normal. The change involved a transformation of r to z (here distinguished by the subscript r) as follows:

$$z_r = 1.1512925 \left(\log \frac{1 + r}{1 - r} \right)$$

The standard deviation of the sampling distribution of any given r is

$$\sigma_{z_r} = 1/\sqrt{N - 3}$$

When these formulas are applied to the problem of Example 14.1, page 331, where $r = 0.683$, the resulting measures of reliability are $z_r = 0.835$ and $\sigma_{z_r} = 0.378$. Thus there is a high probability that the true value of z_r lies between the limits

$$z_r = 0.835 \pm 2 \times 0.378$$

or between 0.079 and 1.591, and it is almost certain that the true value of z is within the three standard deviation limits or between -0.299 and 1.969. These z values may be reduced back to r's and the sampling distribution of r indicated as follows:

σ_{z_r}	z_r	r
-3	-0.299	-0.29
-2	0.079	0.08
-1	0.457	0.43
0	0.835	0.68
1	1.213	0.84
2	1.591	0.92
3	1.969	0.96

Fisher next attacked the problem from an entirely different angle, namely, by estimating the probability of obtaining particular values of r from samples of uncorrelated data. For example, he calculated that an r of 0.63 or more would be likely to occur only once in 20 times, and an r of 0.76 or more only once in 100 times, in drawings of samples of 10 pairs of items having no correlation. Such figures are called

the 5 per cent and 1 per cent levels of probability, respectively. They vary with the size of the sample and the type of correlation but can be tabulated so as to cover the ordinary range of correlation work. These 5 per cent and 1 per cent limits have been selected as the least significant and least highly significant limits, respectively, on the basis of experience chiefly with biological data. The standards may, of course, be modified in accordance with experience in other fields.

The generalized z.—Fisher next broadened the application of the principle just discussed by means of the following formulas:

$$\frac{\Sigma t^2}{n_1} = \Sigma(Y' - \overline{Y})^2 \div n_1 = \frac{\Sigma t^2}{m - 1}$$

$$\frac{\Sigma d^2}{n_2} = \Sigma(Y - Y')^2 \div n_2 = \frac{\Sigma d^2}{N - m}$$

where Y' is T, Y is an item in the dependent series, $\overline{Y}$ is M_Y, and n_1 and n_2 are the degrees of freedom $m - 1$ and $N - m$, respectively (m is the number of series correlated, or number of constants in the regression equation). In simple correlation $n_1 = 1$ and $n_2 = N - 2$. The value of z (not to be confused with z_r) is found as [1]

$$z = 0.5\left(\log_e \frac{\Sigma t^2}{n_1} - \log_e \frac{\Sigma d^2}{n_2}\right)$$

where subtracting the logs has the effect of dividing the variances, and multiplying by 0.5 has the effect of taking their square roots. For convenience the formulas may be written in terms of an ordinary log table as

$$z = 1.1512925 \log\left(\frac{\Sigma t^2}{\Sigma d^2} \times \frac{n_2}{n_1}\right) = 1.1512925 \log\left(\frac{\Sigma t^2}{\Sigma d^2} \times \frac{N - m}{m - 1}\right)$$

It should be noted that the ratio of Σt^2 to Σd^2 is the same as the ratio of r^2 to $1 - r^2$, and the same relationship is true for other correlation measures such as R^2, etc.

The values of z thus obtained may be compared with significant values of z as calculated by Fisher. However, it is more convenient to calculate the statistic F, which omits the log and its coefficient. That is

$$F = \frac{\Sigma t^2}{\Sigma d^2} \times \frac{n_2}{n_1} = \frac{r^2}{1 - r^2} \times \frac{N - m}{m - 1}$$

which may be compared with least significant and least highly significant values of F as read from the table. Other measures of correlation such as R^2 may be substituted for r^2 in these formulas.

It will be seen that F may be interpreted as the ratio of two variances, namely the explained variance and the unexplained variance, each adjusted for its appropriate degrees of freedom. The usual measure of correlation is most commonly calculated in terms of the explained variance and the variance to be explained. Sometimes, the

---

[1] For a more exact statement see Fisher and Yates, *Statistical Tables for Biological Agricultural and Medical Research*, p. 37. Also see preceding pages for 20 per cent, 5 per cent, 1 per cent, and 0.1 per cent levels of z and variance ratios (F).

resulting measure has been corrected for sampling errors by dividing Σd^2 by its degrees of freedom $(N - m)$ and Σy^2 by its degrees of freedom $(N - 1)$, or by correcting r^2 so as to effect the same adjustment. There are theoretical objections to this procedure, and it is no longer necessary, since tables of F and z are both more convenient and more generally applicable.

Finally, it should be emphasized that the sampling distribution of F assumes that the sample from which the measure is derived is drawn from a normal universe, the latter being a continuous rather than a discrete series. If, therefore, the actual universe is extremely abnormal, or if items are rounded so extensively as to destroy the essential feature of continuity, the test of significance may be thereby invalidated.

The gamma function.—The factorial series 1!, 2!, 3!, 4!, etc. (=1, 2, 6, 24, etc.) is discrete, but interpolations may be made by means of a table of the *gamma function* (Γ) as given in *Glover's Tables* and *Pearson's Tables*. This table gives log gamma for numbers from 1 to 2, at intervals of 0.001. The values of gamma thus indicated are the interpolated values of factorial 0! = 1 to 1! = 1. A gamma (not log gamma) table abbreviated is as follows:

X	Γ	X	Γ	X	Γ
1.00	1.0000	1.35	0.8912	1.70	0.9086
1.05	0.9735	1.40	0.8873	1.75	0.9191
1.10	0.9514	1.45	0.8857	1.80	0.9314
1.15	0.9330	1.50	0.8862	1.85	0.9456
1.20	0.9182	1.55	0.8889	1.90	0.9618
1.25	0.9064	1.60	0.8935	1.95	0.9799
1.30	0.8975	1.65	0.9001	2.00	1.0000

The gamma of a tabulated X, as $X = 1.50$, is written $\Gamma 1.50 = 0.8862$, and this is the interpolated factorial of $X - 1$. That is, $0.5! = \Gamma 1.5 = 0.8862$. The interpolation of larger positive factorials depends on the relationship

$$\Gamma X = (X - 1)\Gamma(X - 1) = (X - 1)!$$

or

$$\Gamma(X + 1) = X\Gamma X = X!$$

Hence a large factorial described by $W + f$, where W is a whole number and f a fraction, reduces to

$$(W + f)! = (W + f)(W + f - 1)(W + f - 2)\cdots(1 + f)\Gamma(1 + f)$$

For example:

$$4.5! = 4.5 \times 3.5 \times 2.5 \times 1.5 \times \Gamma 1.5$$
$$= 4.5 \times 3.5 \times 2.5 \times 1.5 \times 0.8862 = 52.34$$

which is between the values 4! = 24 and 5! = 120. For more exact values see the table of log Γ cited above. Logs are given to facilitate the multiplications required for large factorials.

Since $X\Gamma X = \Gamma(X + 1)$, it follows that $\Gamma X = \Gamma(X + 1) \div X$. Hence factorials from -1 to 0, or $-f$, may readily be found as $-f! = \Gamma(1 - f) = [\Gamma(2 - f)] \div (1 - f)$. Thus $-0.25! = \Gamma(0.75) = (\Gamma 1.75) \div 0.75 = 0.9191 \div 0.75 = 1.2254$.

An illustration of the use of the gamma function appears in Example A13.

Example A13

Binomial Curve Fitting, with Interpolations

$$(p + q)^n; \quad y = \frac{n!}{(n - m)! \, m!} p^{n-m} q^m$$

Interpolation by gamma function: $G(x + 1) = xGx = x!$

Let $p = 0.6$ and $q = 0.4$; $n = 4$.

$n - m$	m	p^{n-m}	q^m	$n! \div \overline{(n - m)! \, m!}$	y
5	-1	0.0778	2.5000	$24 \div \infty$	0
$4\frac{1}{2}$	$-\frac{1}{2}$	0.1004	1.5811	$24 \div (52.343 \times 1.772)$	0.0411
4	0	0.1296	1.0000	$24 \div (24 \times 1)$	0.1296
$3\frac{1}{2}$	$\frac{1}{2}$	0.1673	0.6325	$24 \div (11.632 \times 0.886)$	0.2464
3	1	0.2160	0.4000	$24 \div (6 \times 1)$	0.3456
$2\frac{1}{2}$	$1\frac{1}{2}$	0.2789	0.2530	$24 \div (3.323 \times 1.329)$	0.3835
2	2	0.3600	0.1600	$24 \div (2 \times 2)$	0.3456
$1\frac{1}{2}$	$2\frac{1}{2}$	0.4648	0.1012	$24 \div (1.329 \times 3.323)$	0.2556
1	3	0.6000	0.0640	$24 \div (1 \times 6)$	0.1536
$\frac{1}{2}$	$3\frac{1}{2}$	0.7746	0.0405	$24 \div (0.886 \times 11.632)$	0.0731
0	4	1.0000	0.0256	$24 \div (1 \times 24)$	0.0256
$-\frac{1}{2}$	$4\frac{1}{2}$	1.1291	0.0162	$24 \div (1.772 \times 52.343)$	0.0047
-1	5	1.6667	0.1024	$24 \div \infty$	0.0

The Poisson series.—A discontinuous series sometimes applicable to data expressing very small probabilities is known as the Poisson series. According to Fisher this is the most important discontinuous form of distribution, but its uses in business statistics have not as yet been developed to any degree. The frequencies of the series are expressed in terms of the mean as follows:

$$X = 0, 1, 2, 3, 4, \cdots$$

$$f = \left(1, M, \frac{M^2}{2}, \frac{M^3}{2 \times 3}, \frac{M^4}{2 \times 3 \times 4}, \cdots \right)$$

The denominator of each term of the series after the first is X factorial, commonly written $X!$ When the class magnitudes (X) are written in the series indicated above, the sum of the frequencies equals e^M, that is, 2.71828^M. It is a peculiarity of the series that the variance (σ^2) is equal to the mean.

The type of problem in which this dispersion is indicated may be illustrated as follows: assume that in a large number of factories utilizing a certain type of machine, the average accident rate among the operators is 2 per 1,000 per year. If the factories are classified according to their accident rates in a given year their frequencies might be expected to approximate the Poisson type of distribution. The calculation of this distribution appears in Example A14, where X is the accident rate in individual fac-

tories (decimals omitted), and f is the number of factories reporting each specified rate. The frequencies are calculated from the known mean ($M = 2$) according to the formula given above. The mean and variance have been calculated from X and f in order to check their identity with M of the formula.

EXAMPLE A14

THE POISSON SERIES

Data: an assumed distribution in which $M = 2$.

$$X = 0, 1, 2, 3, \cdots$$

$$f = 1, \frac{M}{1}, \frac{M^2}{1 \times 2}, \frac{M^3}{1 \times 2 \times 3}, \cdots\cdots, \text{ in general, } \frac{M^X}{X!}$$

X	f	Xf	X^2f
0	1.	0.	0.
1	2.	2.	2.
2	2.	4.	8.
3	1.333333	4.	12.
4	0.666667	2.666667	10.666667
5	0.266667	1.333333	6.666667
6	0.088889	0.533333	3.200000
7	0.025397	0.177778	1.244444
8	0.006349	0.050794	0.406349
9	0.001411	0.012698	0.114286
10	0.000282	0.002822	0.028219
11	0.000051	0.000564	0.006208
12	0.000009	0.000103	0.001231
13	0.000001	0.000017	0.000222
14		0.000003	0.000037
15			0.000006

Totals: $N = 7.389055$ $\Sigma X = 14.778109$ $\Sigma X^2 = 44.334336$

$e^M = 7.389056$ $M = 2.000000$ $\dfrac{\Sigma X^2}{N} = 6.000000$

Correction: $M^2 = 4.000000$

$\sigma^2 = 2.000000$

MISCELLANEOUS NOTES

Reliability tables and charts.—In preceding chapters reference has been made to tables by which reliability in various types of problems may be estimated. In Figs. A3, 4, and 5, such tables are presented in graphic form. The authors are indebted to R. A. Fisher and George W. Snedecor for this material.

It should be noted that, in agricultural statistics, analysis has proceeded to a point where complex measurements of reliability are highly important and indeed essential to many phases of the work. The science of business statistics has not yet reached such a degree of complexity, but it is rapidly approaching a stage where the more complex measures of reliability are required, and it is hoped that the accompanying charts will serve as a contribution to that end.
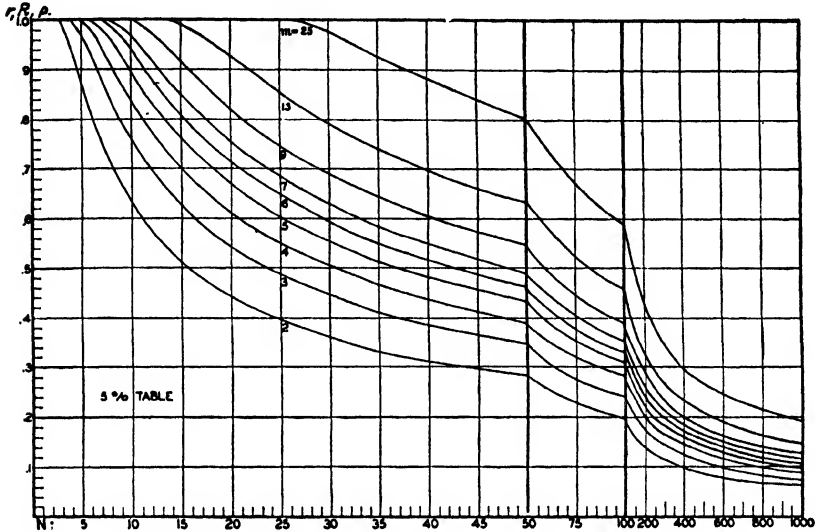
FIG. A3.—The Least Significant Value of r, R, etc. A value as great or greater would appear only once in 20 times by chance. To use the chart:

1. Locate the given number of pairs of items, N, on the horizontal scale at the base of the chart.

2. On the curve representing the given number of variables (constants in the regression equation) designated as m, locate a point directly above the N previously noted (the point of intersection of the curve and the ordinate of the N).

3. Read the appropriate measure of correlation on the vertical scale directly to the left of this point. For example, if N = 31 and m = 3, the measure of correlation that may be exceeded once in 20 times by mere chance is 0.44, which is the height at which the curve m = 3 crosses the ordinate of 31. If the computed measure (not corrected for sampling) is smaller than that read from the chart, it is not considered significant (see H. A. Wallace and G. W. Snedecor, op. cit.).

The table of F (page 586) may be substituted for Figs. A3 and A4. For r, R, etc.,

$$F = \frac{r^2}{1 - r^2} \times \frac{N - m}{m - 1}$$
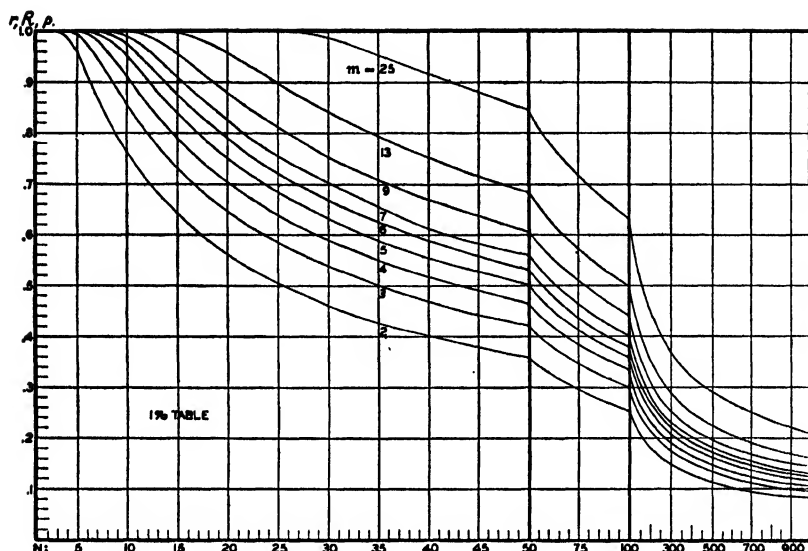
FIG. A4.—The Least Highly Significant Value of r, R, etc. A value as great or greater would appear only once in 100 times by chance. To use the chart:

1. Locate the given number of pairs of items, N, on the horizontal scale at the base of the chart.

2. On the curve representing the given number of variables (constants in the regression equation) designated as m, locate a point directly above the N previously noted (the point of intersection of the curve and the ordinate of N).

3. Read the appropriate measure of correlation on the vertical scale directly to the left of this point of intersection. For example, if $N = 31$ and $m = 3$, the measure of correlation that may be exceeded once in 100 times by pure chance is 0.53, which is the height at which the ordinate of N intersects the curve, $m = 3$. If the computed measure (not corrected for sampling) is greater than that thus discovered from the chart, it is considered highly significant (see Wallace and Snedecor, *op. cit.*).
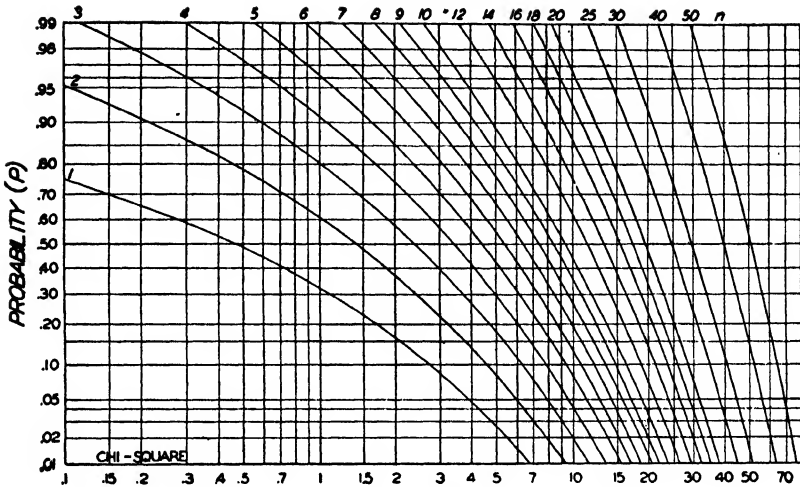
FIG. A5.—The Chi-Square Test of Significance. To use the chart:

1. Locate the computed value of chi square on the horizontal or base scale.

2. Select the appropriate curve for the degrees of freedom, $n = (N - m)$.

3. Note the point of intersection of the ordinate from the point first noted and the selected curve.

4. Read the measure of significance on the vertical scale directly to the left of this point of intersection. The value thus determined, considered as a percentage, represents the number of times in 100 samples that a greater variability from the assumed normal might appear by mere chance. Variation may be regarded as significant at 5 per cent and highly significant at 1 per cent or less.

For higher degrees of freedom (n) compute $\sqrt{2\chi^2} - \sqrt{2n - 1}$ and interpret the result as χ/σ in a table of the normal curve.

SUPPLEMENTARY TABLES

LOGARITHMS

The use of logarithms.—Every number has a corresponding logarithm, or log, which may be obtained from a table of logarithms. Conversely, if the log of a number is given, the number (antilog) may also be obtained from a table.

A logarithm consists of two parts, an integer (positive or negative) usually called the characteristic, and a fraction in decimal form usually called the mantissa. The significance of these two parts will appear later. For example:

LOGARITHM	INTEGER OR CHARACTERISTIC	FRACTION OR MANTISSA
2.8686	2	0.8686
0.8686 − 3	−3	0.8686

The logarithms of numbers that are powers of 10 may be written without the use of a table. For example:

NUMBER	LOG
100	2
10	1
1	0
0.1	−1
0.01	−2

This example indicates that a log is in theory the exponent which, applied to 10, equates the number. On the basis of this theory, the various uses of logarithms in calculations may be explained.

A. *To find the log of a given number.*

1. Place a mark (as subscript x) immediately *after* the first significant figure of the given number and note the number of places (positive or negative) between this mark and the decimal point. This is the log integer. Examples:

GIVEN NUMBER	NUMBER, MARKED	CHARACTERISTIC
7390	7$_x$390	3
739	7$_x$39	2
7.39	7$_x$39	0
0.0739	0.07$_x$39	−2
0.00739	0.007$_x$39	−3

2. Disregarding the position of the decimal point, look up the given number in the margins of a log table, and write the corresponding log as given in the body of the table, prefixing a decimal point. This is the log fraction, or mantissa. Example: the mantissa of 7390; 739; 0.0739; etc., is .8686.

3. Combine the characteristic and mantissa. Positive characteristics precede the fraction; negative characteristics follow. Examples:

GIVEN NUMBER	LOG
7390	3.8686
739	2.8686
7.39	0.8686
0.0739	0.8686 − 2
0.00739	0.8686 − 3

In order to make the negative characteristics uniform in any problem, they are often written as a combination of positive and negative characteristics. However, in statistical work it is usually more convenient to write them as first indicated, except in the case of a log that is to be divided by a certain figure. In this case the negative integer should be this figure, or a multiple of it, and a positive integer should be prefixed to balance any change that may thus be made. Examples:

Log	Divisor	Log Rewritten	Log Divided
$0.8686 - 1$	2	$1.8686 - 2$	$0.9343 - 1$
$0.8686 - 4$	3	$2.8686 - 6$	$0.9562 - 2$

In some cases, as when the divisor consists of several figures, or when other complex calculations are to be made, the log with a negative integer should be reduced by subtraction to a simple negative log. Thus $0.8686 - 1 = -0.1314$; $0.8686 - 2 = -1.1314$; etc. In this case the final result should be changed back by subtraction to the usual form.

B. *To find the antilog of a given log.*

1. Disregarding the characteristic of the log, look up the mantissa in the body of a log table, and, from the margins, note the number corresponding to it. This is the antilog, irrespective of the position of the decimal point. Thus, given the logarithm 0.8686, the antilog figures are found to be 739, the position of the decimal point being undetermined.

2. Place a mark (as subscript x) after the first significant figure of the antilog figures thus found. Point off decimally to the right (positive) or left (negative) as many places as are indicated by the characteristic, prefixing or annexing as many ciphers as may be necessary. Example:

Log	Antilog Figures	Antilog
3.8686	7_x39	7390
0.8686	7_x39	7.39
$0.8686 - 2$	7_x39	0.0739

Note: In finding the log or antilog by the foregoing method, the mark (as subscript x) following the first significant figure of the antilog may, of course, be omitted, provided that the position which it is used to mark is mentally noted. The mark merely indicates the position of the decimal point as it is understood to be placed in the margins of the tables.

A graphic table of logarithms.—The four-place graphic table of logarithms and antilogarithms (pages 566–571) is reprinted from Lacroix and Ragot, *A Graphic Table Combining Logarithms and Anti-Logarithms*, by permission of the publishers, The Macmillan Company, New York. The first digit of the number of which the logarithm is to be taken is read in the column headed N, and succeeding figures are read in the numbers and subdivisions on the upper edge of the scale until the required point is located. The required logarithm is similarly read from the column headed L to the numbers and subdivisions below the scale, at the required point. Antilogarithms may be found by reversing this process. The rules regarding decimals and characteristics apply as before. With care, results may be read to five places. The student would do well to obtain the full five-place table in the reference cited, as it is perhaps the most convenient and accurate table available.

TABLE OF LOGARITHMS

No.	0	1	2	3	4	5	6	7	8	9
1.0	0.0000	0.0043	0.0086	0.0128	0.0170	0.0212	0.0253	0.0294	0.0334	0.0374
1.1	.0414	.0453	.0492	.0531	.0569	.0607	.0645	.0682	.0719	.0755
1.2	.0792	.0828	.0864	.0899	.0934	.0969	.1004	.1038	.1072	.1106
1.3	.1139	.1173	.1206	.1239	.1271	.1303	.1335	.1367	.1399	.1430
1.4	.1461	.1492	.1523	.1553	.1584	.1614	.1644	.1673	.1703	.1732
1.5	.1761	.1790	.1818	.1847	.1875	.1903	.1931	.1959	.1987	.2014
1.6	.2041	.2068	.2095	.2122	.2148	.2175	.2201	.2227	.2253	.2279
1.7	.2304	.2330	.2355	.2380	.2405	.2430	.2455	.2480	.2504	.2529
1.8	.2553	.2577	.2601	.2625	.2648	.2672	.2695	.2718	.2742	.2765
1.9	.2788	.2810	.2833	.2856	.2878	.2900	.2923	.2945	.2967	.2989
2.0	.3010	.3032	.3054	.3075	.3096	.3118	.3139	.3160	.3181	.3201
2.1	.3222	.3243	.3263	.3284	.3304	.3324	.3345	.3365	.3385	.3404
2.2	.3424	.3444	.3464	.3483	.3502	.3522	.3541	.3560	.3579	.3598
2.3	.3617	.3636	.3655	.3674	.3692	.3711	.3729	.3747	.3766	.3784
2.4	.3802	.3820	.3838	.3856	.3874	.3892	.3909	.3927	.3945	.3962
2.5	.3979	.3997	.4014	.4031	.4048	.4065	.4082	.4099	.4116	.4133
2.6	.4150	.4166	.4183	.4200	.4216	.4232	.4249	.4265	.4281	.4298
2.7	.4314	.4330	.4346	.4362	.4378	.4393	.4409	.4425	.4440	.4456
2.8	.4472	.4487	.4502	.4518	.4533	.4548	.4564	.4579	.4594	.4609
2.9	.4624	.4639	.4654	.4669	.4683	.4698	.4713	.4728	.4742	.4757
3.0	.4771	.4786	.4800	.4814	.4829	.4843	.4857	.4871	.4886	.4900
3.1	.4914	.4928	.4942	.4955	.4969	.4983	.4997	.5011	.5024	.5038
3.2	.5051	.5065	.5079	.5092	.5105	.5119	.5132	.5145	.5159	.5172
3.3	.5185	.5198	.5211	.5224	.5237	.5250	.5263	.5276	.5289	.5302
3.4	.5315	.5328	.5340	.5353	.5366	.5378	.5391	.5403	.5416	.5428
3.5	.5441	.5453	.5465	.5478	.5490	.5502	.5514	.5527	.5539	.5551
3.6	.5563	.5575	.5587	.5599	.5611	.5623	.5635	.5647	.5658	.5670
3.7	.5682	.5694	.5705	.5717	.5729	.5740	.5752	.5763	.5775	.5786
3.8	.5798	.5809	.5821	.5832	.5843	.5855	.5866	.5877	.5888	.5899
3.9	.5911	.5922	.5933	.5944	.5955	.5966	.5977	.5988	.5999	.6010
4.0	.6021	.6031	.6042	.6053	.6064	.6075	.6085	.6096	.6107	.6117
4.1	.6128	.6138	.6149	.6160	.6170	.6180	.6191	.6201	.6212	.6222
4.2	.6232	.6243	.6253	.6263	.6274	.6284	.6294	.6304	.6314	.6325
4.3	.6335	.6345	.6355	.6365	.6375	.6385	.6395	.6405	.6415	.6425
4.4	.6435	.6444	.6454	.6464	.6474	.6484	.6493	.6503	.6513	.6522
4.5	.6532	.6542	.6551	.6561	.6571	.6580	.6590	.6599	.6609	.6618
4.6	.6628	.6637	.6646	.6656	.6665	.6675	.6684	.6693	.6702	.6712
4.7	.6721	.6730	.6739	.6749	.6758	.6767	.6776	.6785	.6794	.6803
4.8	.6812	.6821	.6830	.6839	.6848	.6857	.6866	.6875	.6884	.6893
4.9	.6902	.6911	.6920	.6928	.6937	.6946	.6955	.6964	.6972	.6981
5.0	.6990	.6998	.7007	.7016	.7024	.7033	.7042	.7050	.7059	.7067
5.1	.7076	.7084	.7093	.7101	.7110	.7118	.7126	.7135	.7143	.7152
5.2	.7160	.7168	.7177	.7185	.7193	.7202	.7210	.7218	.7226	.7235
5.3	.7243	.7251	.7259	.7267	.7275	.7284	.7292	.7300	.7308	.7316
5.4	.7324	.7332	.7340	.7348	.7356	.7364	.7372	.7380	.7388	.7396

TABLE OF LOGARITHMS—*Continued*

No.	0	1	2	3	4	5	6	7	8	9
5.5	0.7404	0.7412	0.7419	0.7427	0.7435	0.7443	0.7451	0.7459	0.7466	0.7474
5.6	.7482	.7490	.7497	.7505	.7513	.7520	.7528	.7536	.7543	.7551
5.7	.7559	.7566	.7574	.7582	.7589	.7597	.7604	.7612	7619	.7627
5.8	.7634	.7642	.7649	.7657	.7664	.7672	.7679	.7686	.7694	.7701
5.9	.7709	.7716	.7723	.7731	.7738	.7745	.7752	.7760	.7767	.7774
6.0	.7782	.7789	.7796	.7803	.7810	.7818	.7825	.7832	.7839	.7846
6.1	.7853	.7860	.7868	.7875	.7882	.7889	.7896	.7903	.7910	.7917
6.2	.7924	.7931	.7938	.7945	.7952	.7959	.7966	.7973	.7980	.7987
6.3	.7993	.8000	.8007	.8014	.8021	.8028	.8035	.8041	.8048	.8055
6.4	.8062	.8069	.8075	.8082	.8089	.8096	.8102	.8109	.8116	.8122
6.5	.8129	.8136	.8142	.8149	.8156	.8162	.8169	.8176	.8182	.8189
6.6	.8195	.8202	.8209	.8215	.8222	.8228	.8235	.8241	.8248	.8254
6.7	.8261	.8267	.8274	.8280	.8287	.8293	.8299	.8306	.8312	.8319
6.8	.8325	.8331	.8338	.8344	.8351	.8357	.8363	.8370	.8376	.8382
6.9	8388	.8395	.8401	.8407	.8414	.8420	.8426	.8432	.8439	.8445
7.0	.8451	.8457	.8463	.8470	.8476	.8482	.8488	.8494	.8500	.8506
7.1	.8513	.8519	.8525	.8531	.8537	.8543	.8549	.8555	.8561	.8567
7.2	.8573	.8579	.8585	.8591	.8597	.8603	.8609	.8615	.8621	.8627
7.3	.8633	.8639	.8645	.8651	.8657	.8663	.8669	.8675	.8681	.8686
7.4	.8692	.8698	.8704	.8710	.8716	.8722	.8727	.8733	.8739	.8745
7.5	.8751	.8756	.8762	.8768	.8774	.8779	.8785	.8791	.8797	.8802
7.6	.8808	.8814	.8820	.8825	.8831	.8837	.8842	.8848	.8854	.8859
7.7	.8865	.8871	.8876	.8882	.8887	.8893	.8899	.8904	.8910	.8915
7.8	.8921	.8927	.8932	.8938	.8943	.8949	.8954	.8960	.8965	.8971
7.9	.8976	.8982	.8987	.8993	.8998	.9004	.9009	.9015	.9020	.9025
8.0	.9031	.9036	.9042	.9047	.9053	.9058	.9063	.9069	.9074	.9079
8.1	.9085	.9090	.9096	.9101	.9106	.9112	.9117	.9122	.9128	.9133
8.2	.9138	.9143	.9149	.9154	.9159	.9165	.9170	.9175	.9180	.9186
8.3	.9191	.9196	.9201	.9206	.9212	.9217	.9222	.9227	.9232	.9238
8.4	.9243	.9248	.9253	.9258	.9263	.9269	.9274	.9279	.9284	.9289
8.5	.9294	.9299	.9304	.9309	.9315	.9320	.9325	.9330	.9335	.9340
8.6	.9345	.9350	.9355	.9360	.9365	.9370	.9375	.9380	.9385	.9390
8.7	.9395	.9400	.9405	.9410	.9415	.9420	.9425	.9430	.9435	.9440
8.8	.9445	.9450	.9455	.9460	.9465	.9469	.9474	.9479	.9484	.9489
8.9	.9494	.9499	.9504	.9509	.9513	.9518	.9523	.9528	.9533	.9538
9.0	.9542	.9547	.9552	.9557	.9562	.9566	.9571	.9576	.9581	.9586
9.1	.9590	.9595	.9600	.9605	.9609	.9614	.9619	.9624	.9628	.9633
9.2	.9638	.9643	.9647	.9652	.9657	.9661	.9666	.9671	.9675	.9680
9.3	.9685	.9689	.9694	.9699	.9703	.9708	.9713	.9717	.9722	.9727
9.4	.9731	.9736	.9741	.9745	.9750	.9754	.9759	.9763	.9768	.9773
9.5	.9777	.9782	.9786	.9791	.9795	.9800	.9805	.9809	.9814	.9818
9.6	.9823	.9827	.9832	.9836	.9841	.9845	.9850	.9854	.9859	.9863
9.7	.9868	.9872	.9877	.9881	.9886	.9890	.9894	.9899	.9903	.9908
9.8	.9912	.9917	.9921	.9926	.9930	.9934	.9939	.9943	.9948	.9952
9.9	.9956	.9961	.9965	.9969	.9974	.9978	.9983	.9987	.9991	.9996

| N | L | N 1000 – 1560   L .0000 – .1932 |

1

0

00 .... 01 02
00 01 02 03 04 05 06 07 08 09 10

03 .... 04 05
11 12 13 14 15 16 17 18 19 20 21

05 .... 06 07
22 23 24 25 26 27 28 29 30 31

08 .... 09 10
32 33 34 35 36 37 38 39 40 41

10 .... 11 12
42 43 44 45 46 47 48 49 50 51

13 .... 14 15
52 53 54 55 56 57 58 59 60

15 .... 16 17
61 62 63 64 65 66 67 68 69 70

18 .... 19 20
70 71 72 73 74 75 76 77 78 79

20 .... 21 .... 22 23
80 81 82 83 84 85 86 87 88 89 90

23 .... 24 .... 25 26
90 91 92 93 94 95 96 97 98 00

1

26 .... 27 .... 28 29
01 02 03 04 05 06 07 08 09 10

29 .... 30 .... 31 32
11 12 13 14 15 16 17 18 19 20

32 .... 33 .... 34 35
21 22 23 24 25 26 27 28 29 30

35 .... 36 .... 37 38
31 32 33 34 35 36 37 38 39

38 .... 39 .... 40 41
40 41 42 43 44 45 46 47 48 49

41 .... 42 .... 43 44
50 51 52 53 54 55 56 57 58

44 .... 45 .... 46 47
59 60 61 62 63 64 65 66 67

47 .... 48 .... 49 50
68 69 70 71 72 73 74 75 76

50 .... 51 .... 52 53
76 77 78 79 80 81 82 83 84

53 .... 54 .... 55 56
85 86 87 88 89 90 91 92 93

N	L	N 1560 – 2480	L .1931 – .3945

N	L			
2		**N 2480 – 4000**		**L .3944 – .6021**

N	L	N 4000 – 6000	L .6020 –.7782

4		00 01 02 03 04 05 06 07 08 09 10
	6	02 03 04 05 06 07 08 09 10 11 12
		10 11 12 13 14 15 16 17 18 19 20
		13 14 15 16 17 18 19 20 21 22 23
		20 21 22 23 24 25 26 27 28 29 30
		24 25 26 27 28 29 30 31 32 33
		30 31 32 33 34 35 36 37 38 39 40
		34 35 36 37 38 39 40 41 42 43
		40 41 42 43 44 45 46 47 48 49 50
		44 45 46 47 48 49 50 51 52 53
		50 51 52 53 54 55 56 57 58 59 60
		54 55 56 57 58 59 60 61 62
		60 61 62 63 64 65 66 67 68 69 70
		63 64 65 66 67 68 69 70 71 72
		70 71 72 73 74 75 76 77 78 79 80
		72 73 74 75 76 77 78 79 80 81
		80 81 82 83 84 85 86 87 88 89 90
		82 83 84 85 86 87 88 89 90
		90 91 92 93 94 95 96 97 98 99 00
		91 92 93 94 95 96 97 98 99
5		00 01 02 03 04 05 06 07 08 09 10
		99 00 01 02 03 04 05 06 07
		10 11 12 13 14 15 16 17 18 19 20
	7	08 09 10 11 12 13 14 15 16
		20 21 22 23 24 25 26 27 28 29 30
		16 17 18 19 20 21 22 23 24
		30 31 32 33 34 35 36 37 38 39 40
		25 26 27 28 29 30 31 32
		40 41 42 43 44 45 46 47 48 49 50
		33 34 35 36 37 .38 39 40
		50 51 52 53 54 55 56 57 58 59 60
		41 42 43 44 45 46 47 48
		60 61 62 63 64 65 66 67 68 69 70
		49 50 51 52 53 54 55
		70 71 72 73 74 75 76 77 78 79 80
		56 57 58 59 60 61 62 63
		80 81 82 83 84 85 86 87 88 89 90
		64 65 66 67 68 69 70
		90 91 92 93 94 95 96 97 98 99 00
		71 72 73 74 75 76 77 78

N	L	N 6000 – 8000	L .7781 – .9031

6

7
00 01 02 03 04 05 06 07 08 09 10
79 80 81 82 83 84 85

10 11 12 13 14 15 16 17 18 19 20
86 87 88 89 90 91 92

20 21 22 23 24 25 26 27 28 29 30
93 94 95 96 97 98 99

30 31 32 33 34 35 36 37 38 39 40
00 01 02 03 04 05 06

40 41 42 43 44 45 46 47 48 49 50

8
07 08 09 10 11 12 13

50 51 52 53 54 55 56 57 58 59 60
13 14 15 16 17 18 19

60 61 62 63 64 65 66 67 68 69 70
20 21 22 23 24 25 26

70 71 72 73 74 75 76 77 78 79 80
26 27 28 29 30 31 32

80 81 82 83 84 85 86 87 88 89 90
33 34 35 36 37 38

90 91 92 93 94 95 96 97 98 99 00
39 40 41 42 43 44 45

7
00 01 02 03 04 05 06 07 08 09 10
45 46 47 48 49 50 51

10 11 12 13 14 15 16 17 18 19 20
52 53 54 55 56 57

20 21 22 23 24 25 26 27 28 29 30
58 59 60 61 62 63

30 31 32 33 34 35 36 37 38 39 40
64 65 66 67 68 69

40 41 42 43 44 45 46 47 48 49 50
70 71 72 73 74 75

50 51 52 53 54 55 56 57 58 59 60
75 76 77 78 79 80

60 61 62 63 64 65 66 67 68 69 70
81 82 83 84 85 86

70 71 72 73 74 75 76 77 78 79 80
87 88 89 90 91 92

80 81 82 83 84 85 86 87 88 89 90
92 93 94 95 96 97

90 91 92 93 94 95 96 97 98 99 00
98 99 00 01 02 03

N	L	N 8000 – 10000	L .9030 – .0000

8

9 03 04 05 06 07 08

00 01 02 03 04 05 06 07 08 09 10

10 11 12 13 14 15 16 17 18 19 20
09 10 11 12 13

20 21 22 23 24 25 26 27 28 29 30
14 15 16 17 18 19

30 31 32 33 34 35 36 37 38 39 40
19 20 21 22 23 24

40 41 42 43 44 45 46 47 48 49 50
25 26 27 28 29

50 51 52 53 54 55 56 57 58 59 60
30 31 32 33 34

60 61 62 63 64 65 66 67 68 69 70
35 36 37 38 39

70 71 72 73 74 75 76 77 78 79 80
40 41 42 43 44

80 81 82 83 84 85 86 87 88 89 90
45 46 47 48 49

90 91 92 93 94 95 96 97 98 99 00
50 51 52 53 54

9 00 01 02 03 04 05 06 07 08 09 10
55 56 57 58 59

10 11 12 13 14 15 16 17 18 19 20
59 60 61 62 63

20 21 22 23 24 25 26 27 28 29 30
64 65 66 67 68

30 31 32 33 34 35 36 37 38 39 40
69 70 71 72 73

40 41 42 43 44 45 46 47 48 49 50
74 75 76 77

50 51 52 53 54 55 56 57 58 59 60
78 79 80 81 82

60 61 62 63 64 65 66 67 68 69 70
83 84 85 86

70 71 72 73 74 75 76 77 78 79 80
87 88 89 90 91

80 81 82 83 84 85 86 87 88 89 90
92 93 94 95

90 91 92 93 94 95 96 97 98 99 00
96 97 98 99 00

SQUARES, SQUARE ROOTS, AND RECIPROCALS TO 1000

No.	Square	Square Root	Reciprocal × 100	No.	Square	Square Root	Reciprocal × 100
1	1	1.0000000	100.0000000	51	26 01	7.1414284	1.9607843
2	4	1.4142136	50.0000000	52	27 04	7.2111026	1.9230769
3	9	1.7320508	33.3333333	53	28 09	7.2801099	1.8867925
4	16	2.0000000	25.0000000	54	29 16	7.3484692	1.8518519
5	25	2.2360680	20.0000000	55	30 25	7.4161985	1.8181818
6	36	2.4494897	16.6666667	56	31 36	7.4833148	1.7857143
7	49	2.6457513	14.2857143	57	32 49	7.5498344	1.7543860
8	64	2.8284271	12.5000000	58	33 64	7.6157731	1.7241379
9	81	3.0000000	11.1111111	59	34 81	7.6811457	1.6949153
10	1 00	3.1622777	10.0000000	60	36 00	7.7459667	1.6666667
11	1 21	3.3166248	9.0909091	61	37 21	7.8102497	1.6393443
12	1 44	3.4641016	8.3333333	62	38 44	7.8740079	1.6129032
13	1 69	3.6055513	7.6923077	63	39 69	7.9372539	1.5873016
14	1 96	3.7416574	7.1428571	64	40 96	8.0000000	1.5625000
15	2 25	3.8729833	6.6666667	65	42 25	8.0622577	1.5384615
16	2 56	4.0000000	6.2500000	66	43 56	8.1240384	1.5151515
17	2 89	4.1231056	5.8823529	67	44 89	8.1853528	1.4925373
18	3 24	4.2426407	5.5555556	68	46 24	8.2462113	1.4705882
19	3 61	4.3588989	5.2631579	69	47 61	8.3066239	1.4492754
20	4 00	4.4721360	5.0000000	70	49 00	8.3666003	1.4285714
21	4 41	4.5825757	4.7619048	71	50 41	8.4261498	1.4084507
22	4 84	4.6904158	4.5454545	72	51 84	8.4852814	1.3888889
23	5 29	4.7958315	4.3478261	73	53 29	8.5440037	1.3698630
24	5 76	4.8989795	4.1666667	74	54 76	8.6023253	1.3513514
25	6 25	5.0000000	4.0000000	75	56 25	8.6602540	1.3333333
26	6 76	5.0990195	3.8461538	76	57 76	8.7177979	1.3157895
27	7 29	5.1961524	3.7037037	77	59 29	8.7749644	1.2987013
28	7 84	5.2915026	3.5714286	78	60 84	8.8317609	1.2820513
29	8 41	5.3851648	3.4482759	79	62 41	8.8881944	1.2658228
30	9 00	5.4772256	3.3333333	80	64 00	8.9442719	1.2500000
31	9 61	5.5677644	3.2258065	81	65 61	9.0000000	1.2345679
32	10 24	5.6568542	3.1250000	82	67 24	9.0553851	1.2195122
33	10 89	5.7445626	3.0303030	83	68 89	9.1104336	1.2048193
34	11 56	5.8309519	2.9411765	84	70 56	9.1651514	1.1904762
35	12 25	5.9160798	2.8571429	85	72 25	9.2195445	1.1764706
36	12 96	6.0000000	2.7777778	86	73 96	9.2736185	1.1627907
37	13 69	6.0827625	2.7027027	87	75 69	9.3273791	1.1494253
38	14 44	6.1644140	2.6315789	88	77 44	9.3808315	1.1363636
39	15 21	6.2449980	2.5641026	89	79 21	9.4339811	1.1235955
40	16 00	6.3245553	2.5000000	90	81 00	9.4868330	1.1111111
41	16 81	6.4031242	2.4390244	91	82 81	9.5393920	1.0989011
42	17 64	6.4807407	2.3809524	92	84 64	9.5916630	1.0869565
43	18 49	6.5574385	2.3255814	93	86 49	9.6436508	1.0752688
44	19 36	6.6332496	2.2727273	94	88 36	9.6953597	1.0638298
45	20 25	6.7082039	2.2222222	95	90 25	9.7467943	1.0526316
46	21 16	6.7823300	2.1739130	96	92 16	9.7979590	1.0416667
47	22 09	6.8556546	2.1276596	97	94 09	9.8488578	1.0309278
48	23 04	6.9282032	2.0833333	98	96 04	9.8994949	1.0204082
49	24 01	7.0000000	2.0408163	99	98 01	9.9498744	1.0101010
50	25 00	7.0710678	2.0000000	100	1 00 00	10.0000000	1.0000000

No.	Square	Square Root	Reciprocal × 10⁹	No.	Square	Square Root	Reciprocal × 10⁹
101	1 02 01	10.0498756	9900990	151	2 28 01	12.2882057	6622517
102	1 04 04	10.0995049	9803922	152	2 31 04	12.3288280	6578947
103	1 06 09	10.1488916	9708738	153	2 34 09	12.3693169	6535948
104	1 08 16	10.1980390	9615385	154	2 37 16	12.4096736	6493506
105	1 10 25	10.2469508	9523810	155	2 40 25	12.4498996	6451613
106	1 12 36	10.2956301	9433962	156	2 43 36	12.4899960	6410256
107	1 14 49	10.3440804	9345794	157	2 46 49	12.5299641	6369427
108	1 16 64	10.3923048	9259259	158	2 49 64	12.5698051	6329114
109	1 18 81	10.4403065	9174312	159	2 52 81	12.6095202	6289308
110	1 21 00	10.4880885	9090909	160	2 56 00	12.6491106	6250000
111	1 23 21	10.5356538	9009009	161	2 59 21	12.6885775	6211180
112	1 25 44	10.5830052	8928571	162	2 62 44	12.7279221	6172840
113	1 27 69	10.6301458	8849558	163	2 65 69	12.7671453	6134969
114	1 29 96	10.6770783	8771930	164	2 68 96	12.8062485	6097561
115	1 32 25	10.7238053	8695652	165	2 72 25	12.8452326	6060606
116	1 34 56	10.7703296	8620690	166	2 75 56	12.8840987	6024096
117	1 36 89	10.8166538	8547009	167	2 78 89	12.9228480	5988024
118	1 39 24	10.8627805	8474576	168	2 82 24	12.9614814	5952381
119	1 41 61	10.9087121	8403361	169	2 85 61	13.0000000	5917160
120	1 44 00	10.9544512	8333333	170	2 89 00	13.0384048	5882353
121	1 46 41	11.0000000	8264463	171	2 92 41	13.0766968	5847953
122	1 48 84	11.0453610	8196721	172	2 95 84	13.1148770	5813953
123	1 51 29	11.0905365	8130081	173	2 99 29	13.1529464	5780347
124	1 53 76	11.1355287	8064516	174	3 02 76	13.1909060	5747126
125	1 56 25	11.1803399	8000000	175	3 06 25	13.2287566	5714286
126	1 58 76	11.2249722	7936508	176	3 09 76	13.2664992	5681818
127	1 61 29	11.2694277	7874016	177	3 13 29	13.3041347	5649718
128	1 63 84	11.3137085	7812500	178	3 16 84	13.3416641	5617978
129	1 66 41	11.3578167	7751938	179	3 20 41	13.3790882	5586592
130	1 69 00	11.4017543	7692308	180	3 24 00	13.4164079	5555556
131	1 71 61	11.4455231	7633588	181	3 27 61	13.4536240	5524862
132	1 74 24	11.4891253	7575758	182	3 31 24	13.4907376	5494505
133	1 76 89	11.5325626	7518797	183	3 34 89	13.5277493	5464481
134	1 79 56	11.5758369	7462687	184	3 38 56	13.5646600	5434783
135	1 82 25	11.6189500	7407407	185	3 42 25	13.6014705	5405405
136	1 84 96	11.6619038	7352941	186	3 45 96	13.6381817	5376344
137	1 87 69	11.7046999	7299270	187	3 49 69	13.6747943	5347594
138	1 90 44	11.7473401	7246377	188	3 53 44	13.7113092	5319149
139	1 93 21	11.7898261	7194245	189	3 57 21	13.7477271	5291005
140	1 96 00	11.8321596	7142857	190	3 61 00	13.7840488	5263158
141	1 98 81	11.8743421	7092199	191	3 64 81	13.8202750	5235602
142	2 01 64	11.9163753	7042254	192	3 68 64	13.8564065	5208333
143	2 04 49	11.9582607	6993007	193	3 72 49	13.8924440	5181347
144	2 07 36	12.0000000	6944444	194	3 76 36	13.9283883	5154639
145	2 10 25	12.0415946	6896552	195	3 80 25	13.9642400	5128205
146	2 13 16	12.0830460	6849315	196	3 84 16	14.0000000	5102041
147	2 16 09	12.1243557	6802721	197	3 88 09	14.0356688	5076142
148	2 19 04	12.1655251	6756757	198	3 92 04	14.0712473	5050505
149	2 22 01	12.2065556	6711409	199	3 96 01	14.1067360	5025126
150	2 25 00	12.2474487	6666667	200	4 00 00	14.1421356	5000000

No.	Square	Square Root	Reciprocal × 10⁹	No.	Square	Square Root	Reciprocal × 10⁹
201	4 04 01	14.1774469	4975124	251	6 30 01	15.8429795	3984064
202	4 08 04	14.2126704	4950495	252	6 35 04	15.8745079	3968254
203	4 12 09	14.2478068	4926108	253	6 40 09	15.9059737	3952569
204	4 16 16	14.2828569	4901961	254	6 45 16	15.9373775	3937008
205	4 20 25	14.3178211	4878049	255	6 50 25	15.9687194	3921569
206	4 24 36	14.3527001	4854369	256	6 55 36	16.0000000	3906250
207	4 28 49	14.3874946	4830918	257	6 60 49	16.0312195	3891051
208	4 32 64	14.4222051	4807692	258	6 65 64	16.0623784	3875969
209	4 36 81	14.4568323	4784689	259	6 70 81	16.0934769	3861004
210	4 41 00	14.4913767	4761905	260	6 76 00	16.1245155	3846154
211	4 45 21	14.5258390	4739336	261	6 81 21	16.1554944	3831418
212	4 49 44	14.5602198	4716981	262	6 86 44	16.1864141	3816794
213	4 53 69	14.5945195	4694836	263	6 91 69	16.2172747	3802281
214	4 57 96	14.6287388	4672897	264	6 96 96	16.2480768	3787879
215	4 62 25	14.6628783	4651163	265	7 02 25	16.2788206	3773585
216	4 66 56	14.6969385	4629630	266	7 07 56	16.3095064	3759398
217	4 70 89	14.7309199	4608295	267	7 12 89	16.3401346	3745318
218	4 75 24	14.7648231	4587156	268	7 18 24	16.3707055	3731343
219	4 79 61	14.7986486	4566210	269	7 23 61	16.4012195	3717472
220	4 84 00	14.8323970	4545455	270	7 29 00	16.4316767	3703704
221	4 88 41	14.8660687	4524887	271	7 34 41	16.4620776	3690037
222	4 92 84	14.8996644	4504505	272	7 39 84	16.4924225	3676471
223	4 97 29	14.9331845	4484305	273	7 45 29	16.5227116	3663004
224	5 01 76	14.9666295	4464286	274	7 50 76	16.5529454	3649635
225	5 06 25	15.0000000	4444444	275	7 56 25	16.5831240	3636364
226	5 10 76	15.0332964	4424779	276	7 61 76	16.6132477	3623188
227	5 15 29	15.0665192	4405286	277	7 67 29	16.6433170	3610108
228	5 19 84	15.0996689	4385965	278	7 72 84	16.6733320	3597122
229	5 24 41	15.1327460	4366812	279	7 78 41	16.7032931	3584229
230	5 29 00	15.1657509	4347826	280	7 84 00	16.7332005	3571429
231	5 33 61	15.1986842	4329004	281	7 89 61	16.7630546	3558719
232	5 38 24	15.2315462	4310345	282	7 95 24	16.7928556	3546099
233	5 42 89	15.2643375	4291845	283	8 00 89	16.8226038	3533569
234	5 47 56	15.2970585	4273504	284	8 06 56	16.8522995	3521127
235	5 52 25	15.3297097	4255319	285	8 12 25	16.8819430	3508772
236	5 56 96	15.3622915	4237288	286	8 17 96	16.9115345	3496503
237	5 61 69	15.3948043	4219409	287	8 23 69	16.9410743	3484321
238	5 66 44	15.4272486	4201681	288	8 29 44	16.9705627	3472222
239	5 71 21	15.4596248	4184100	289	8 35 21	17.0000000	3460208
240	5 76 00	15.4919334	4166667	290	8 41 00	17.0293864	3448276
241	5 80 81	15.5241747	4149378	291	8 46 81	17.0587221	3436426
242	5 85 64	15.5563492	4132231	292	8 52 64	17.0880075	3424658
243	5 90 49	15.5884573	4115226	293	8 58 49	17.1172428	3412969
244	5 95 36	15.6204994	4098361	294	8 64 36	17.1464282	3401361
245	6 00 25	15.6524758	4081633	295	8 70 25	17.1755640	3389831
246	6 05 16	15.6843871	4065041	296	8 76 16	17.2046505	3378378
247	6 10 09	15.7162336	4048583	297	8 82 09	17.2336879	3367003
248	6 15 04	15.7480157	4032258	298	8 88 04	17.2626765	3355705
249	6 20 01	15.7797338	4016064	299	8 94 01	17.2916165	3344482
250	6 25 00	15.8113883	4000000	300	9 00 00	17.3205081	3333333

No.	Square	Square Root	Recipro-cal $\times 10^9$	No.	Square	Square Root	Recipro-cal $\times 10^9$
301	9 06 01	17.3493516	3322259	351	12 32 01	18.7349940	2849003
302	9 12 04	17.3781472	3311258	352	12 39 04	18.7616630	2840909
303	9 18 09	17.4068952	3300330	353	12 46 09	18.7882942	2832861
304	9 24 16	17.4355958	3289474	354	12 53 16	18.8148877	2824859
305	9 30 25	17.4642492	3278689	355	12 60 25	18.8414437	2816901
306	9 36 36	17.4928557	3267974	356	12 67 36	18.8679623	2808989
307	9 42 49	17.5214155	3257329	357	12 74 49	18.8944436	2801120
308	9 48 64	17.5499288	3246753	358	12 81 64	18.9208879	2793296
309	9 54 81	17.5783958	3236246	359	12 88 81	18.9472953	2785515
310	9 61 00	17.6068169	3225806	360	12 96 00	18.9736660	2777778
311	9 67 21	17.6351921	3215434	361	13 03 21	19.0000000	2770083
312	9 73 44	17.6635217	3205128	362	13 10 44	19.0262976	2762431
313	9 79 69	17.6918060	3194888	363	13 17 69	19.0525589	2754821
314	9 85 96	17.7200451	3184713	364	13 24 96	19.0787840	2747253
315	9 92 25	17.7482393	3174603	365	13 32 25	19.1049732	2739726
316	9 98 56	17.7763888	3164557	366	13 39 56	19.1311265	2732240
317	10 04 89	17.8044928	3154574	367	13 46 89	19.1572441	2724796
318	10 11 24	17.8325545	3144654	368	13 54 24	19.1833261	2717391
319	10 17 61	17.8605711	3134796	369	13 61 61	19.2093727	2710027
320	10 24 00	17.8885438	3125000	370	13 69 00	19.2353841	2702703
321	10 30 41	17.9164729	3115265	371	13 76 41	19.2613603	2695418
322	10 36 84	17.9443584	3105590	372	13 83 84	19.2873015	2688172
323	10 43 29	17.9722008	3095975	373	13 91 29	19.3132079	2680965
324	10 49 76	18.0000000	3086420	374	13 98 76	19.3390796	2673797
325	10 56 25	18.0277564	3076923	375	14 06 25	19.3649167	2666667
326	10 62 76	18.0554701	3067485	376	14 13 76	19.3907194	2659574
327	10 69 29	18.0831413	3058104	377	14 21 29	19.4164878	2652520
328	10 75 84	18.1107703	3048780	378	14 28 84	19.4422221	2645503
329	10 82 41	18.1383571	3039514	379	14 36 41	19.4679223	2638522
330	10 89 00	18.1659021	3030303	380	14 44 00	19.4935887	2631579
331	10 95 61	18.1934054	3021148	381	14 51 61	19.5192213	2624672
332	11 02 24	18.2208672	3012048	382	14 59 24	19.5448203	2617801
333	11 08 89	18.2482876	3003003	383	14 66 89	19.5703858	2610966
334	11 15 56	18.2756669	2994012	384	14 74 56	19.5959179	2604167
335	11 22 25	18.3030052	2985075	385	14 82 25	19.6214169	2597403
336	11 28 96	18.3303028	2976190	386	14 89 96	19.6468827	2590674
337	11 35 69	18.3575598	2967359	387	14 97 69	19.6723156	2583979
338	11 42 44	18.3847763	2958580	388	15 05 44	19.6977156	2577320
339	11 49 21	18.4119526	2949853	389	15 13 21	19.7230829	2570694
340	11 56 00	18.4390889	2941176	390	15 21 00	19.7484177	2564103
341	11 62 81	18.4661853	2932551	391	15 28 81	19.7737199	2557545
342	11 69 64	18.4932420	2923977	392	15 36 64	19.7989899	2551020
343	11 76 49	18.5202592	2915452	393	15 44 49	19.8242276	2544529
344	11 83 36	18.5472370	2906977	394	15 52 36	19.8494332	2538071
345	11 90 25	18.5741756	2898551	395	15 60 25	19.8746069	2531646
346	11 97 16	18.6010752	2890173	396	15 68 16	19.8997487	2525253
347	12 04 09	18.6279360	2881844	397	15 76 09	19.9248588	2518892
348	12 11 04	18.6547581	2873563	398	15 84 04	19.9499373	2512563
349	12 18 01	18.6815417	2865330	399	15 92 01	19.9749844	2506266
350	12 25 00	18.7082869	2857143	400	16 00 00	20.0000000	2500000

No.	Square	Square Root	Recipro-cal × 10⁹	No.	Square	Square Root	Recipro-cal × 10⁹
401	16 08 01	20.0249844	2493766	451	20 34 01	21.2367606	2217295
402	16 16 04	20.0499377	2487562	452	20 43 04	21.2602916	2212389
403	16 24 09	20.0748599	2481390	453	20 52 09	21.2837967	2207506
404	16 32 16	20.0997512	2475248	454	20 61 16	21.3072758	2202643
405	16 40 25	20.1246118	2469136	455	20 70 25	21.3307290	2197802
406	16 48 36	20.1494417	2463054	456	20 79 36	21.3541565	2192982
407	16 56 49	20.1742410	2457002	457	20 88 49	21.3775583	2188184
408	16 64 64	20.1990099	2450980	458	20 97 64	21.4009346	2183406
409	16 72 81	20.2237484	2444988	459	21 06 81	21.4242853	2178649
410	16 81 00	20.2484567	2439024	460	21 16 00	21.4476106	2173913
411	16 89 21	20.2731349	2433090	461	21 25 21	21.4709106	2169197
412	16 97 44	20.2977831	2427184	462	21 34 44	21.4941853	2164502
413	17 05 69	20.3224014	2421308	463	21 43 69	21.5174348	2159827
414	17 13 96	20.3469899	2415459	464	21 52 96	21.5406592	2155172
415	17 22 25	20.3715488	2409639	465	21 62 25	21.5638587	2150538
416	17 30 56	20.3960781	2403846	466	21 71 56	21.5870331	2145923
417	17 38 89	20.4205779	2398082	467	21 80 89	21.6101828	2141328
418	17 47 24	20.4450483	2392344	468	21 90 24	21.6333077	2136752
419	17 55 61	20.4694895	2386635	469	21 99 61	21.6564078	2132196
420	17 64 00	20.4939015	2380952	470	22 09 00	21.6794834	2127660
421	17 72 41	20.5182845	2375297	471	22 18 41	21.7025344	2123142
422	17 80 84	20.5426386	2369668	472	22 27 84	21.7255610	2118644
423	17 89 29	20.5669638	2364066	473	22 37 29	21.7485632	2114165
424	17 97 76	20.5912603	2358491	474	22 46 76	21.7715411	2109705
425	18 06 25	20.6155281	2352941	475	22 56 25	21.7944947	2105263
426	18 14 76	20.6397674	2347418	476	22 65 76	21.8174242	2100840
427	18 23 29	20.6639783	2341920	477	22 75 29	21.8403297	2096436
428	18 31 84	20.6881609	2336449	478	22 84 84	21.8632111	2092050
429	18 40 41	20.7123152	2331002	479	22 94 41	21.8860686	2087683
430	18 49 00	20.7364414	2325581	480	23 04 00	21.9089023	2083333
431	18 57 61	20.7605395	2320186	481	23 13 61	21.9317122	2079002
432	18 66 24	20.7846097	2314815	482	23 23 24	21.9544984	2074689
433	18 74 89	20.8086520	2309469	483	23 32 89	21.9772610	2070393
434	18 83 56	20.8326667	2304147	484	23 42 56	22.0000000	2066116
435	18 92 25	20.8566536	2298851	485	23 52 25	22.0227155	2061856
436	19 00 96	20.8806130	2293578	486	23 61 96	22.0454077	2057613
437	19 09 69	20.9045450	2288330	487	23 71 69	22.0680765	2053388
438	19 18 44	20.9284495	2283105	488	23 81 44	22.0907220	2049180
439	19 27 21	20.9523268	2277904	489	23 91 21	22.1133444	2044990
440	19 36 00	20.9761770	2272727	490	24 01 00	22.1359436	2040816
441	19 44 81	21.0000000	2267574	491	24 10 81	22.1585198	2036660
442	19 53 64	21.0237960	2262443	492	24 20 64	22.1810730	2032520
443	19 62 49	21.0475652	2257336	493	24 30 49	22.2036033	2028398
444	19 71 36	21.0713075	2252252	494	24 40 36	22.2261108	2024291
445	19 80 25	21.0950231	2247191	495	24 50 25	22.2485955	2020202
446	19 89 16	21.1187121	2242152	496	24 60 16	22.2710575	2016129
447	19 98 09	21.1423745	2237136	497	24 70 09	22.2934968	2012072
448	20 07 04	21.1660105	2232143	498	24 80 04	22.3159136	2008032
449	20 16 01	21.1896201	2227171	499	24 90 01	22.3383079	2004008
450	20 25 00	21.2132034	2222222	500	25 00 00	22.3606798	2000000

No.	Square	Square Root	Recipro-cal × 10⁹	No.	Square	Square Root	Recipro-cal × 10⁹
501	25 10 01	22.3830293	1996008	551	30 36 01	23.4733892	1814882
502	25 20 04	22.4053565	1992032	552	30 47 04	23.4946802	1811594
503	25 30 09	22.4276615	1988072	553	30 58 09	23.5159520	1808318
504	25 40 16	22.4499443	1984127	554	30 69 16	23.5372046	1805054
505	25 50 25	22.4722051	1980198	555	30 80 25	23.5584380	1801802
506	25 60 36	22.4944438	1976285	556	30 91 36	23.5796522	1798561
507	25 70 49	22.5166605	1972387	557	31 02 49	23.6008474	1795332
508	25 80 64	22.5388553	1968504	558	31 13 64	23.6220236	1792115
509	25 90 81	22.5610283	1964637	559	31 24 81	23.6431808	1788909
510	26 01 00	22.5831796	1960784	560	31 36 00	23.6643191	1785714
511	26 11 21	22.6053091	1956947	561	31 47 21	23.6854386	1782531
512	26 21 44	22.6274170	1953125	562	31 58 44	23.7065392	1779359
513	26 31 69	22.6495033	1949318	563	31 69 69	23.7276210	1776199
514	26 41 96	22.6715681	1945525	564	31 80 96	23.7486842	1773050
515	26 52 25	22.6936114	1941748	565	31 92 25	23.7697286	1769912
516	26 62 56	22.7156334	1937984	566	32 03 56	23.7907545	1766784
517	26 72 89	22.7376340	1934236	567	32 14 89	23.8117618	1763668
518	26 83 24	22.7596134	1930502	568	32 26 24	23.8327506	1760563
519	26 93 61	22.7815715	1926782	569	32 37 61	23.8537209	1757469
520	27 04 00	22.8035085	1923077	570	32 49 00	23.8746728	1754386
521	27 14 41	22.8254244	1919386	571	32 60 41	23.8956063	1751313
522	27 24 84	22.8473193	1915709	572	32 71 84	23.9165215	1748252
523	27 35 29	22.8691933	1912046	573	32 83 29	23.9374184	1745201
524	27 45 76	22.8910463	1908397	574	32 94 76	23.9582971	1742160
525	27 56 25	22.9128785	1904762	575	33 06 25	23.9791576	1739130
526	27 66 76	22.9346899	1901141	576	33 17 76	24.0000000	1736111
527	27 77 29	22.9564806	1897533	577	33 29 29	24.0208243	1733102
528	27 87 84	22.9782506	1893939	578	33 40 84	24.0416306	1730104
529	27 98 41	23.0000000	1890359	579	33 52 41	24.0624188	1727116
530	28 09 00	23.0217289	1886792	580	33 64 00	24.0831892	1724138
531	28 19 61	23.0434372	1883239	581	33 75 61	24.1039416	1721170
532	28 30 24	23.0651252	1879699	582	33 87 24	24.1246762	1718213
533	28 40 89	23.0867928	1876173	583	33 98 89	24.1453929	1715266
534	28 51 56	23.1084400	1872659	584	34 10 56	24.1660919	1712329
535	28 62 25	23.1300670	1869159	585	34 22 25	24.1867732	1709402
536	28 72 96	23.1516738	1865672	586	34 33 96	24.2074369	1706485
537	28 83 69	23.1732605	1862197	587	34 45 69	24.2280829	1703578
538	28 94 44	23.1948270	1858736	588	34 57 44	24.2487113	1700680
539	29 05 21	23.2163735	1855288	589	34 69 21	24.2693222	1697793
540	29 16 00	23.2379001	1851852	590	34 81 00	24.2899156	1694915
541	29 26 81	23.2594067	1848429	591	34 92 81	24.3104916	1692047
542	29 37 64	23.2808935	1845018	592	35 04 64	24.3310501	1689189
543	29 48 49	23.3023604	1841621	593	35 16 49	24.3515913	1686341
544	29 59 36	23.3238076	1838235	594	35 28 36	24.3721152	1683502
545	29 70 25	23.3452351	1834862	595	35 40 25	24.3926218	1680672
546	29 81 16	23.3666429	1831502	596	35 52 16	24.4131112	1677852
547	29 92 09	23.3880311	1828154	597	35 64 09	24.4335834	1675042
548	30 03 04	23.4093998	1824818	598	35 76 04	24.4540385	1672241
549	30 14 01	23.4307490	1821494	599	35 88 01	24.4744765	1669449
550	30 25 00	23.4520788	1818182	600	36 00 00	24.4948974	1666667

No.	Square	Square Root	Recipro-cal $\times 10^9$	No.	Square	Square Root	Recipro-cal $\times 10^9$
601	36 12 01	24.5153013	1663894	651	42 38 01	25.5147016	1536098
602	36 24 04	24.5356883	1661130	652	42 51 04	25.5342907	1533742
603	36 36 09	24.5560583	1658375	653	42 64 09	25.5538647	1531394
604	36 48 16	24.5764115	1655629	654	42 77 16	25.5734237	1529052
605	36 60 25	24.5967478	1652893	655	42 90 25	25.5929678	1526718
606	36 72 36	24.6170673	1650165	656	43 03 36	25.6124969	1524390
607	36 84 49	24.6373700	1647446	657	43 16 49	25.6320112	1522070
608	36 96 64	24.6576560	1644737	658	43 29 64	25.6515107	1519757
609	37 08 81	24.6779254	1642036	659	43 42 81	25.6709953	1517451
610	37 21 00	24.6981781	1639344	660	43 56 00	25.6904652	1515152
611	37 33 21	24.7184142	1636661	661	43 69 21	25.7099203	1512859
612	37 45 44	24.7386338	1633987	662	43 82 44	25.7293607	1510574
613	37 57 69	24.7588368	1631321	663	43 95 69	25.7487864	1508296
614	37 69 96	24.7790234	1628664	664	44 08 96	25.7681975	1506024
615	37 82 25	24.7991935	1626016	665	44 22 25	25.7875939	1503759
616	37 94 56	24.8193473	1623377	666	44 35 56	25.8069758	1501502
617	38 06 89	24.8394847	1620746	667	44 48 89	25.8263431	1499250
618	38 19 24	24.8596058	1618123	668	44 62 24	25.8456960	1497006
619	38 31 61	24.8797106	1615509	669	44 75 61	25.8650343	1494768
620	38 44 00	24.8997992	1612903	670	44 89 00	25.8843582	1492537
621	38 56 41	24.9198716	1610306	671	45 02 41	25.9036677	1490313
622	38 68 84	24.9399278	1607717	672	45 15 84	25.9229628	1488095
623	38 81 29	24.9599679	1605136	673	45 29 29	25.9422435	1485884
624	38 93 76	24.9799920	1602564	674	45 42 76	25.9615100	1483680
625	39 06 25	25.0000000	1600000	675	45 56 25	25.9807621	1481481
626	39 18 76	25.0199920	1597444	676	45 69 76	26.0000000	1479290
627	39 31 29	25.0399681	1594896	677	45 83 29	26.0192237	1477105
628	39 43 84	25.0599282	1592357	678	45 96 84	26.0384331	1474926
629	39 56 41	25.0798724	1589825	679	46 10 41	26.0576284	1472754
630	39 69 00	25.0998008	1587302	680	46 24 00	26.0768096	1470588
631	39 81 61	25.1197134	1584786	681	46 37 61	26.0959767	1468429
632	39 94 24	25.1396102	1582278	682	46 51 24	26.1151297	1466276
633	40 06 89	25.1594913	1579779	683	46 64 89	26.1342687	1464129
634	40 19 56	25.1793566	1577287	684	46 78 56	26.1533937	1461988
635	40 32 25	25.1992063	1574803	685	46 92 25	26.1725047	1459854
636	40 44 96	25.2190404	1572327	686	47 05 96	26.1916017	1457726
637	40 57 69	25.2388589	1569859	687	47 19 69	26.2106848	1455604
638	40 70 44	25.2586619	1567398	688	47 33 44	26.2297541	1453488
639	40 83 21	25.2784493	1564945	689	47 47 21	26.2488095	1451379
640	40 96 00	25.2982213	1562500	690	47 61 00	26.2678511	1449275
641	41 08 81	25.3179778	1560062	691	47 74 81	26.2868789	1447178
642	41 21 64	25.3377189	1557632	692	47 88 64	26.3058929	1445087
643	41 34 49	25.3574447	1555210	693	48 02 49	26.3248932	1443001
644	41 47 36	25.3771551	1552795	694	48 16 36	26.3438797	1440922
645	41 60 25	25.3968502	1550388	695	48 30 25	26.3628527	1438849
646	41 73 16	25.4165301	1547988	696	48 44 16	26.3818119	1436782
647	41 86 09	25.4361947	1545595	697	48 58 09	26.4007576	1434720
648	41 99 04	25.4558441	1543210	698	48 72 04	26.4196896	1432665
649	42 12 01	25.4754784	1540832	699	48 86 01	26.4386081	1430615
650	42 25 00	25.4950976	1538462	700	49 00 00	26.4575131	1428571

No.	Square	Square Root	Recipro-cal $\times 10^9$	No.	Square	Square Root	Recipro-cal $\times 10^9$
701	49 14 01	26.4764046	1426534	751	56 40 01	27.4043792	1331558
702	49 28 04	26.4952826	1424501	752	56 55 04	27.4226184	1329787
703	49 42 09	26.5141472	1422475	753	56 70 09	27.4408455	1328021
704	49 56 16	26.5329983	1420455	754	56 85 16	27.4590604	1326260
705	49 70 25	26.5518361	1418440	755	57 00 25	27.4772633	1324503
706	49 84 36	26.5706605	1416431	756	57 15 36	27.4954542	1322751
707	49 98 49	26.5894716	1414427	757	57 30 49	27.5136330	1321004
708	50 12 64	26.6082694	1412429	758	57 45 64	27.5317998	1319261
709	50 26 81	26.6270539	1410437	759	57 60 81	27.5499546	1317523
710	50 41 00	26.6458252	1408451	760	57 76 00	27.5680975	1315789
711	50 55 21	26.6645833	1406470	761	57 91 21	27.5862284	1314060
712	50 69 44	26.6833281	1404494	762	58 06 44	27.6043475	1312336
713	50 83 69	26.7020598	1402525	763	58 21 69	27.6224546	1310616
714	50 97 96	26.7207784	1400560	764	58 36 96	27.6405499	1308901
715	51 12 25	26.7394839	1398601	765	58 52 25	27.6586334	1307190
716	51 26 56	26.7581763	1396648	766	58 67 56	27.6767050	1305483
717	51 40 89	26.7768557	1394700	767	58 82 89	27.6947648	1303781
718	51 55 24	26.7955220	1392758	768	58 98 24	27.7128129	1302083
719	51 69 61	26.8141754	1390821	769	59 13 61	27.7308492	1300390
720	51 84 00	26.8328157	1388889	770	59 29 00	27.7488739	1298701
721	51 98 41	26.8514432	1386963	771	59 44 41	27.7668868	1297017
722	52 12 84	26.8700577	1385042	772	59 59 84	27.7848880	1295337
723	52 27 29	26.8886593	1383126	773	59 75 29	27.8028775	1293661
724	52 41 76	26.9072481	1381215	774	59 90 76	27.8208555	1291990
725	52 56 25	26.9258240	1379310	775	60 06 25	27.8388218	1290323
726	52 70 76	26.9443872	1377410	776	60 21 76	27.8567766	1288660
727	52 85 29	26.9629375	1375516	777	60 37 29	27.8747197	1287001
728	52 99 84	26.9814751	1373626	778	60 52 84	27.8926514	1285347
729	53 14 41	27.0000000	1371742	779	60 68 41	27.9105715	1283697
730	53 29 00	27.0185122	1369863	780	60 84 00	27.9284801	1282051
731	53 43 61	27.0370117	1367989	781	60 99 61	27.9463772	1280410
732	53 58 24	27.0554985	1366120	782	61 15 24	27.9642629	1278772
733	53 72 89	27.0739727	1364256	783	61 30 89	27.9821372	1277139
734	53 87 56	27.0924344	1362398	784	61 46 56	28.0000000	1275510
735	54 02 25	27.1108834	1360544	785	61 62 25	28.0178515	1273885
736	54 16 96	27.1293199	1358696	786	61 77 96	28.0356915	1272265
737	54 31 69	27.1477439	1356852	787	61 93 69	28.0535203	1270648
738	54 46 44	27.1661554	1355014	788	62 09 44	28.0713377	1269036
739	54 61 21	27.1845544	1353180	789	62 25 21	28.0891438	1267427
740	54 76 00	27.2029410	1351351	790	62 41 00	28.1069386	1265823
741	54 90 81	27.2213152	1349528	791	62 56 81	28.1247222	1264223
742	55 05 64	27.2396769	1347709	792	62 72 64	28.1424946	1262626
743	55 20 49	27.2580263	1345895	793	62 88 49	28.1602557	1261034
744	55 35 36	27.2763634	1344086	794	63 04 36	28.1780056	1259446
745	55 50 25	27.2946881	1342282	795	63 20 25	28.1957444	1257862
746	55 65 16	27.3130006	1340483	796	63 36 16	28.2134720	1256281
747	55 80 09	27.3313007	1338688	797	63 52 09	28.2311884	1254705
748	55 95 04	27.3495887	1336898	798	63 68 04	28.2488938	1253133
749	56 10 01	27.3678644	1325113	799	63 84 01	28.2665881	1251564
750	56 25 00	27.3861279	1333333	800	64 00 00	28.2842712	1250000

No.	Square	Square Root	Recipro-cal $\times 10^9$	No.	Square	Square Root	Recipro-cal $\times 10^9$
801	64 16 01	28.3019434	1248439	851	72 42 01	29.1719043	1175088
802	64 32 04	28.3196045	1246883	852	72 59 04	29.1890390	1173709
803	64 48 09	28.3372546	1245330	853	72 76 09	29.2061637	1172333
804	64 64 16	28.3548938	1243781	854	72 93 16	29.2232784	1170960
805	64 80 25	28.3725219	1242236	855	73 10 25	29.2403830	1169591
806	64 96 36	28.3901391	1240695	856	73 27 36	29.2574777	1168224
807	65 12 49	28.4077454	1239157	857	73 44 49	29.2745623	1166861
808	65 28 64	28.4253408	1237624	858	73 61 64	29.2916370	1165501
809	65 44 81	28.4429253	1236094	859	73 78 81	29.3087018	1164144
810	65 61 00	28.4604989	1234568	860	73 96 00	29.3257566	1162791
811	65 77 21	28.4780617	1233046	861	74 13 21	29.3428015	1161440
812	65 93 44	28.4956137	1231527	862	74 30 44	29.3598365	1160093
813	66 09 69	28.5131549	1230012	863	74 47 69	29.3768616	1158749
814	66 25 96	28.5306852	1228501	864	74 64 96	29.3938769	1157407
815	66 42 25	28.5482048	1226994	865	74 82 25	29.4108823	1156069
816	66 58 56	28.5657137	1225490	866	74 99 56	29.4278779	1154734
817	66 74 89	28.5832119	1223990	867	75 16 89	29.4448637	1153403
818	66 91 24	28.6006993	1222494	868	75 34 24	29.4618397	1152074
819	67 07 61	28.6181760	1221001	869	75 51 61	29.4788059	1150748
820	67 24 00	28.6356421	1219512	870	75 69 00	29.4957624	1149425
821	67 40 41	28.6530976	1218027	871	75 86 41	29.5127091	1148106
822	67 56 84	28.6705424	1216545	872	76 03 84	29.5296461	1146789
823	67 73 29	28.6879766	1215067	873	76 21 29	29.5465734	1145475
824	67 89 76	28.7054002	1213592	874	76 38 76	29.5634910	1144165
825	68 06 25	28.7228132	1212121	875	76 56 25	29.5803989	1142857
826	68 22 76	28.7402157	1210654	876	76 73 76	29.5972972	1141553
827	68 39 29	28.7576077	1209190	877	76 91 29	29.6141858	1140251
828	68 55 84	28.7749891	1207729	878	77 08 84	29.6310648	1138952
829	68 72 41	28.7923601	1206273	879	77 26 41	29.6479342	1137656
830	68 89 00	28.8097206	1204819	880	77 44 00	29.6647939	1136364
831	69 05 61	28.8270706	1203369	881	77 61 61	29.6816442	1135074
832	69 22 24	28.8444102	1201923	882	77 79 24	29.6984848	1133787
833	69 38 89	28.8617394	1200480	883	77 96 89	29.7153159	1132503
834	69 55 56	28.8790582	1199041	884	78 14 56	29.7321375	1131222
835	69 72 25	28.8963666	1197605	885	78 32 25	29.7489496	1129944
836	69 88 96	28.9136646	1196172	886	78 49 96	29.7657521	1128668
837	70 05 69	28.9309523	1194743	887	78 67 69	29.7825452	1127396
838	70 22 44	28.9482297	1193317	888	78 85 44	29.7993289	1126126
839	70 39 21	28.9654967	1191895	889	79 03 21	29.8161030	1124859
840	70 56 00	28.9827535	1190476	890	79 21 00	29.8328678	1123596
841	70 72 81	29.0000000	1189061	891	79 38 81	29.8496231	1122334
842	70 89 64	29.0172363	1187648	892	79 56 64	29.8663690	1121076
843	71 06 49	29.0344623	1186240	893	79 74 49	29.8831056	1119821
844	71 23 36	29.0516781	1184834	894	79 92 36	29.8998328	1118568
845	71 40 25	29.0688837	1183432	895	80 10 25	29.9165506	1117318
846	71 57 16	29.0860791	1182033	896	80 28 16	29.9332591	1116071
847	71 74 09	29.1032644	1180638	897	80 46 09	29.9499583	1114827
848	71 91 04	29.1204396	1179245	898	80 64 04	29.9666481	1113586
849	72 08 01	29.1376046	1177856	899	80 82 01	29.9833287	1112347
850	72 25 00	29.1547595	1176471	900	81 00 00	30.0000000	1111111

No.	Square	Square Root	Recipro- cal $\times 10^9$	No.	Square	Square Root	Recipro- cal $\times 10^9$
901	81 18 01	30.0166620	1109878	951	90 44 01	30.8382879	1051525
902	81 36 04	30.0333148	1108647	952	90 63 04	30.8544972	1050420
903	81 54 09	30.0499584	1107420	953	90 82 09	30.8706981	1049318
904	81 72 16	30.0665928	1106195	954	91 01 16	30.8868904	1048218
905	81 90 25	30.0832179	1104972	955	91 20 25	30.9030743	1047120
906	82 08 36	30.0998339	1103753	956	91 39 36	30.9192497	1046025
907	82 26 49	30.1164407	1102536	957	91 58 49	30.9354166	1044932
908	82 44 64	30.1330383	1101322	958	91 77 64	30.9515751	1043841
909	82 62 81	30.1496269	1100110	959	91 96 81	30.9677251	1042753
910	82 81 00	30.1662063	1098901	960	92 16 00	30.9838668	1041667
911	82 99 21	30.1827765	1097695	961	92 35 21	31.0000000	1040583
912	83 17 44	30.1993377	1096491	962	92 54 44	31.0161248	1039501
913	83 35 69	30.2158899	1095290	963	92 73 69	31.0322413	1038422
914	83 53 96	30.2324329	1094092	964	92 92 96	31.0483494	1037344
915	83 72 25	30.2489669	1092896	965	93 12 25	31.0644491	1036269
916	83 90 56	30.2654919	1091703	966	93 31 56	31.0805405	1035197
917	84 08 89	30.2820079	1090513	967	93 50 89	31.0966236	1034126
918	84 27 24	30.2985148	1089325	968	93 70 24	31.1126984	1033058
919	84 45 61	30.3150128	1088139	969	93 89 61	31.1287648	1031992
920	84 64 00	30.3315018	1086957	970	94 09 00	31.1448230	1030928
921	84 82 41	30.3479818	1085776	971	94 28 41	31.1608729	1029866
922	85 00 84	30.3644529	1084599	972	94 47 84	31.1769145	1028807
923	85 19 29	30.3809151	1083424	973	94 67 29	31.1929479	1027749
924	85 37 76	30.3973683	1082251	974	94 86 76	31.2089731	1026694
925	85 56 25	30.4138127	1081081	975	95 06 25	31.2249900	1025641
926	85 74 76	30.4302481	1079914	976	95 25 76	31.2409987	1024590
927	85 93 29	30.4466747	1078749	977	95 45 29	31.2569992	1023541
928	86 11 84	30.4630924	1077586	978	95 64 84	31.2729915	1022495
929	86 30 41	30.4795013	1076426	979	95 84 41	31.2889757	1021450
930	86 49 00	30.4959014	1075269	980	96 04 00	31.3049517	1020408
931	86 67 61	30.5122926	1074114	981	96 23 61	31.3209195	1019368
932	86 86 24	30.5286750	1072961	982	96 43 24	31.3368792	1018330
933	87 04 89	30.5450487	1071811	983	96 62 89	31.3528308	1017294
934	87 23 56	30.5614136	1070664	984	96 82 56	31.3687743	1016260
935	87 42 25	30.5777697	1069519	985	97 02 25	31.3847097	1015228
936	87 60 96	30.5941171	1068376	986	97 21 96	31.4006369	1014199
937	87 79 69	30.6104557	1067236	987	97 41 69	31.4165561	1013171
938	87 98 44	30.6267857	1066098	988	97 61 44	31.4324673	1012146
939	88 17 21	30.6431069	1064963	989	97 81 21	31.4483704	1011122
940	88 36 00	30.6594194	1063830	990	98 01 00	31.4642654	1010101
941	88 54 81	30.6757233	1062699	991	98 20 81	31.4801525	1009082
942	88 73 64	30.6920185	1061571	992	98 40 64	31.4960315	1008065
943	88 92 49	30.7083051	1060445	993	98 60 49	31.5119025	1007049
944	89 11 36	30.7245830	1059322	994	98 80 36	31.5277655	1006036
945	89 30 25	30.7408523	1058201	995	99 00 25	31.5436206	1005025
946	89 49 16	30.7571130	1057082	996	99 20 16	31.5594677	1004016
947	89 68 09	30.7733651	1055966	997	99 40 09	31.5753068	1003009
948	89 87 04	30.7896086	1054852	998	99 60 04	31.5911380	1002004
949	90 06 01	30.8058436	1053741	999	99 80 01	31.6069613	1001001
950	90 25 00	30.8220700	1052632	1000	1 00 00 00	31.6227766	1000000

TABLE OF THE NORMAL CURVE

Ordinates (z) and Cumulative Area (A) of the Right Half of the Normal
Curve of Distribution of Unit Area

For cumulative of whole curve, read $.5 \pm A$ for $\pm x/\sigma$. Ordinates are represented in terms of the total area as unity.

x/σ	z	A	x/σ	z	A
0.00	0.39894	0.00000	0.50	0.35207	0.19146
0.01	0.39892	0.00399	0.51	0.35029	0.19497
0.02	0.39886	0.00798	0.52	0.34849	0.19847
0.03	0.39876	0.01197	0.53	0.34667	0.20194
0.04	0.39862	0.01595	0.54	0.34482	0.20540
0.05	0.39844	0.01994	0.55	0.34294	0.20884
0.06	0.39822	0.02392	0.56	0.34105	0.21226
0.07	0.39797	0.02790	0.57	0.33912	0.21566
0.08	0.39767	0.03188	0.58	0.33718	0.21904
0.09	0.39733	0.03586	0.59	0.33521	0.22240
0.10	0.39695	0.03983	0.60	0.33322	0.22575
0.11	0.39654	0.04380	0.61	0.33121	0.22907
0.12	0.39603	0.04776	0.62	0.32918	0.23237
0.13	0.39559	0.05172	0.63	0.32713	0.23565
0.14	0.39505	0.05567	0.64	0.32506	0.23891
0.15	0.39448	0.05962	0.65	0.32297	0.24215
0.16	0.39387	0.06356	0.66	0.32086	0.24537
0.17	0.39322	0.06749	0.67	0.31874	0.24857
0.18	0.39253	0.07142	0.68	0.31659	0.25175
0.19	0.39181	0.07535	0.69	0.31443	0.25490
0.20	0.39104	0.07926	0.70	0.31225	0.25804
0.21	0.39024	0.08317	0.71	0.31006	0.26115
0.22	0.38940	0.08706	0.72	0.30785	0.26424
0.23	0.38853	0.09095	0.73	0.30563	0.26730
0.24	0.38762	0.09483	0.74	0.30339	0.27035
0.25	0.38667	0.09871	0.75	0.30114	0.27337
0.26	0.38568	0.10257	0.76	0.29887	0.27637
0.27	0.38466	0.10642	0.77	0.29659	0.27935
0.28	0.38361	0.11026	0.78	0.29431	0.28230
0.29	0.38251	0.11409	0.79	0.29200	0.28524
0.30	0.38139	0.11791	0.80	0.28969	0.28814
0.31	0.38023	0.12172	0.81	0.28737	0.29103
0.32	0.37903	0.12552	0.82	0.28504	0.29389
0.33	0.37780	0.12930	0.83	0.28269	0.29673
0.34	0.37654	0.13307	0.84	0.28034	0.29955
0.35	0.37524	0.13683	0.85	0.27798	0.30234
0.36	0.37391	0.14058	0.86	0.27562	0.30511
0.37	0.37255	0.14431	0.87	0.27324	0.30785
0.38	0.37115	0.14803	0.88	0.27086	0.31057
0.39	0.36973	0.15173	0.89	0.26848	0.31327
0.40	0.36827	0.15542	0.90	0.26609	0.31594
0.41	0.36678	0.15910	0.91	0.26369	0.31859
0.42	0.36526	0.16276	0.92	0.26129	0.32121
0.43	0.36371	0.16640	0.93	0.25888	0.32381
0.44	0.36213	0.17003	0.94	0.25647	0.32639
0.45	0.36053	0.17364	0.95	0.25406	0.32894
0.46	0.35889	0.17724	0.96	0.25164	0.33147
0.47	0.35723	0.18082	0.97	0.24923	0.33398
0.48	0.35553	0.18439	0.98	0.24681	0.33646
0.49	0.35381	0.18793	0.99	0.24439	0.33891

TABLE OF THE NORMAL CURVE—*Continued*

x/σ	z	A	x/σ	z	A
1.00	0.24197	0.34134	1.50	0.12952	0.43319
1.01	0.23955	0.34375	1.51	0.12758	0.43448
1.02	0.23713	0.34614	1.52	0.12566	0.43574
1.03	0.23471	0.34850	1.53	0.12376	0.43699
1.04	0.23230	0.35083	1.54	0.12188	0.43822
1.05	0.22988	0.35314	1.55	0.12001	0.43943
1.06	0.22747	0.35543	1.56	0.11816	0.44062
1.07	0.22506	0.35769	1.57	0.11632	0.44179
1.08	0.22265	0.35993	1.58	0.11450	0.44295
1.09	0.22025	0.36214	1.59	0.11270	0.44408
1.10	0.21785	0.36433	1.60	0.11092	0.44520
1.11	0.21546	0.36650	1.61	0.10915	0.44630
1.12	0.21307	0.36864	1.62	0.10741	0.44738
1.13	0.21069	0.37076	1.63	0.10567	0.44845
1.14	0.20831	0.37286	1.64	0.10396	0.44950
1.15	0.20594	0.37493	1.65	0.10226	0.45053
1.16	0.20357	0.37698	1.66	0.10059	0.45154
1.17	0.20121	0.37900	1.67	0.09893	0.45254
1.18	0.19886	0.38100	1.68	0.09728	0.45352
1.19	0.19652	0.38298	1.69	0.09566	0.45449
1.20	0.19419	0.38493	1.70	0.09405	0.45543
1.21	0.19186	0.38686	1.71	0.09246	0.45637
1.22	0.18954	0.38877	1.72	0.09089	0.45728
1.23	0.18724	0.39065	1.73	0.08933	0.45818
1.24	0.18494	0.39251	1.74	0.08780	0.45907
1.25	0.18265	0.39435	1.75	0.08628	0.45994
1.26	0.18037	0.39617	1.76	0.08478	0.46080
1.27	0.17810	0.39796	1.77	0.08329	0.46164
1.28	0.17585	0.39973	1.78	0.08183	0.46246
1.29	0.17360	0.40147	1.79	0.08038	0.46327
1.30	0.17137	0.40320	1.80	0.07895	0.46407
1.31	0.16915	0.40490	1.81	0.07754	0.46485
1.32	0.16694	0.40658	1.82	0.07614	0.46562
1.33	0.16474	0.40824	1.83	0.07477	0.46638
1.34	0.16256	0.40988	1.84	0.07341	0.46712
1.35	0.16038	0.41149	1.85	0.07206	0.46784
1.36	0.15822	0.41309	1.86	0.07074	0.46856
1.37	0.15608	0.41466	1.87	0.06943	0.46926
1.38	0.15395	0.41621	1.88	0.06814	0.46995
1.39	0.15183	0.41774	1.89	0.06687	0.47062
1.40	0.14973	0.41924	1.90	0.06562	0.47128
1.41	0.14764	0.42073	1.91	0.06438	0.47193
1.42	0.14556	0.42220	1.92	0.06316	0.47257
1.43	0.14350	0.42364	1.93	0.06195	0.47320
1.44	0.14146	0.42507	1.94	0.06077	0.47381
1.45	0.13943	0.42647	1.95	0.05959	0.47441
1.46	0.13742	0.42786	1.96	0.05844	0.47500
1.47	0.13542	0.42922	1.97	0.05730	0.47558
1.48	0.13344	0.43056	1.98	0.05618	0.47615
1.49	0.13147	0.43189	1.99	0.05508	0.47670

TABLE OF THE NORMAL CURVE—*Continued*

x/σ	z	A	x/σ	z	A
2.00	0.05399	0.47725	2.50	0.01753	0.49379
2.01	0.05292	0.47778	2.51	0.01709	0.49396
2.02	0.05186	0.47831	2.52	0.01667	0.49413
2.03	0.05082	0.47882	2.53	0.01625	0.49430
2.04	0.04980	0.47932	2.54	0.01585	0.49446
2.05	0.04879	0.47982	2.55	0.01545	0.49461
2.06	0.04780	0.48030	2.56	0.01506	0.49477
2.07	0.04682	0.48077	2.57	0.01468	0.49492
2.08	0.04586	0.48124	2.58	0.01431	0.49506
2.09	0.04491	0.48169	2.59	0.01394	0.49520
2.10	0.04398	0.48214	2.60	0.01358	0.49534
2.11	0.04307	0.48257	2.61	0.01323	0.49547
2.12	0.04217	0.48300	2.62	0.01289	0.49560
2.13	0.04128	0.48341	2.63	0.01256	0.49573
2.14	0.04041	0.48382	2.64	0.01223	0.49585
2.15	0.03955	0.48422	2.65	0.01191	0.49598
2.16	0.03871	0.48461	2.66	0.01160	0.49609
2.17	0.03788	0.48500	2.67	0.01130	0.49621
2.18	0.03706	0.48537	2.68	0.01100	0.49632
2.19	0.03626	0.48574	2.69	0.01071	0.49643
2.20	0.03547	0.48610	2.70	0.01042	0.49653
2.21	0.03470	0.48645	2.71	0.01014	0.49664
2.22	0.03394	0.48679	2.72	0.00987	0.49674
2.23	0.03319	0.48713	2.73	0.00961	0.49683
2.24	0.03246	0.48745	2.74	0.00935	0.49693
2.25	0.03174	0.48778	2.75	0.00909	0.49702
2.26	0.03103	0.48809	2.76	0.00885	0.49711
2.27	0.03034	0.48840	2.77	0.00861	0.49720
2.28	0.02965	0.48870	2.78	0.00837	0.49728
2.29	0.02898	0.48899	2.79	0.00814	0.49736
2.30	0.02833	0.48928	2.80	0.00792	0.49744
2.31	0.02768	0.48956	2.81	0.00770	0.49752
2.32	0.02705	0.48983	2.82	0.00748	0.49760
2.33	0.02643	0.49010	2.83	0.00727	0.49767
2.34	0.02582	0.49036	2.84	0.00707	0.49774
2.35	0.02522	0.49061	2.85	0.00687	0.49781
2.36	0.02463	0.49086	2.86	0.00668	0.49788
2.37	0.02406	0.49111	2.87	0.00649	0.49795
2.38	0.02349	0.49134	2.88	0.00631	0.49801
2.39	0.02294	0.49158	2.89	0.00613	0.49807
2.40	0.02239	0.49180	2.90	0.00595	0.49813
2.41	0.02186	0.49202	2.91	0.00578	0.49819
2.42	0.02134	0.49224	2.92	0.00562	0.49825
2.43	0.02083	0.49245	2.93	0.00545	0.49831
2.44	0.02033	0.49266	2.94	0.00530	0.49836
2.45	0.01984	0.49286	2.95	0.00514	0.49841
2.46	0.01936	0.49305	2.96	0.00499	0.49846
2.47	0.01889	0.49324	2.97	0.00485	0.49851
2.48	0.01842	0.49343	2.98	0.00471	0.49856
2.49	0.01797	0.49361	2.99	0.00457	0.49861

TABLE OF THE NORMAL CURVE—*Continued*

x/σ	z	A	x/σ	z	A
3.00	0.00443	0.49865	3.50	0.00087	0.49977
3.01	0.00430	0.49869	3.51	0.00084	0.49978
3.02	0.00417	0.49874	3.52	0.00081	0.49978
3.03	0.00405	0.49878	3.53	0.00079	0.49979
3.04	0.00393	0.49882	3.54	0.00076	0.49980
3.05	0.00381	0.49886	3.55	0.00073	0.49981
3.06	0.00370	0.49889	3.56	0.00071	0.49981
3.07	0.00358	0.49893	3.57	0.00068	0.49982
3.08	0.00348	0.49897	3.58	0.00066	0.49983
3.09	0.00337	0.49900	3.59	0.00063	0.49983
3.10	0.00327	0.49903	3.60	0.00061	0.49984
3.11	0.00317	0.49906	3.61	0.00059	0.49985
3.12	0.00307	0.49910	3.62	0.00057	0.49985
3.13	0.00298	0.49913	3.63	0.00055	0.49986
3.14	0.00288	0.49916	3.64	0.00053	0.49986
3.15	0.00279	0.49918	3.65	0.00051	0.49987
3.16	0.00271	0.49921	3.66	0.00049	0.49987
3.17	0.00262	0.49924	3.67	0.00047	0.49988
3.18	0.00254	0.49926	3.68	0.00046	0.49988
3.19	0.00246	0.49929	3.69	0.00044	0.49989
3.20	0.00238	0.49931	3.70	0.00042	0.49989
3.21	0.00231	0.49934	3.71	0.00041	0.49990
3.22	0.00224	0.49936	3.72	0.00039	0.49990
3.23	0.00216	0.49938	3.73	0.00038	0.49990
3.24	0.00210	0.49940	3.74	0.00037	0.49991
3.25	0.00203	0.49942	3.75	0.00035	0.49991
3.26	0.00196	0.49944	3.76	0.00034	0.49992
3.27	0.00190	0.49946	3.77	0.00033	0.49992
3.28	0.00184	0.49948	3.78	0.00031	0.49992
3.29	0.00178	0.49950	3.79	0.00030	0.49992
3.30	0.00172	0.49952	3.80	0.00029	0.49993
3.31	0.00167	0.49953	3.81	0.00028	0.49993
3.32	0.00161	0.49955	3.82	0.00027	0.49993
3.33	0.00156	0.49957	3.83	0.00026	0.49994
3.34	0.00151	0.49958	3.84	0.00025	0.49994
3.35	0.00146	0.49960	3.85	0.00024	0.49994
3.36	0.00141	0.49961	3.86	0.00023	0.49994
3.37	0.00136	0.49962	3.87	0.00022	0.49995
3.38	0.00132	0.49964	3.88*	0.00021	0.49995
3.39	0.00127	0.49965	3.90	0.00020	0.49995
3.40	0.00123	0.49966	3.91	0.00019	0.49995
3.41	0.00119	0.49968	3.92	0.00018	0.49996
3.42	0.00115	0.49969	3.94	0.00017	0.49996
3.43	0.00111	0.49970	3.95	0.00016	0.49996
3.44	0.00107	0.49971	3.97	0.00015	0.49996
3.45	0.00104	0.49972	3.98	0.00014	0.49997
3.46	0.00100	0.49973	4.00	0.00013	0.49997
3.47	0.00097	0.49974	4.02	0.00012	0.49997
3.48	0.00094	0.49975	4.04	0.00011	0.49997
3.49	0.00090	0.49976	4.06	0.00011	0.49998

* For skipped x/σ items below, read values next preceding.

TABLE OF 5% & 1% F FOR DESIGNATED DEGREES OF FREEDOM IN GREATER AND SMALLER MEAN SQUARE

Reprinted by permission from Davenport and Ekas, *Statistical Methods*, Fourth Edition, John Wiley and Sons, Inc., New York, and from G. W. Snedecor, *Analysis of Variance*, Collegiate Press, Ames, Iowa.

Degrees of Freedom for Smaller Mean Square	Degrees of Freedom for Greater Mean Square														Values of t
	1	2	3	4	5	6	7	8	10	12	16	24	50	∞	
1	161.45	199.50	215.72	224.57	230.17	233.97	236.75	238.89	242.00	243.91	246.50	249.04	251.80	254.32	12.706
	4052.10	4999.03	5403.49	5625.14	5764.08	5859.39	5928.00	5981.34	6056.00	6105.83	6169.00	6234.16	6302.00	6366.48	63.657
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.39	19.41	19.43	19.45	19.47	19.50	4.303
	98.49	99.01	99.17	99.25	99.30	99.33	99.34	99.36	99.40	99.42	99.44	99.46	99.48	99.50	9.925
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.78	8.74	8.69	8.64	8.58	8.53	3.182
	34.12	30.81	29.46	28.71	28.24	27.91	27.67	27.49	27.23	27.05	26.83	26.60	26.35	26.12	5.841
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.84	5.77	5.70	5.63	2.776
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.54	14.37	14.15	13.93	13.69	13.46	4.604
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.60	4.53	4.44	4.36	2.571
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.05	9.89	9.68	9.47	9.24	9.02	4.032
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.92	3.84	3.75	3.67	2.447
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.87	7.72	7.52	7.31	7.09	6.88	3.707
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.63	3.57	3.49	3.41	3.32	3.23	2.365
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.62	6.47	6.27	6.07	5.85	5.65	3.499
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.34	3.28	3.20	3.12	3.03	2.93	2.306
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.82	5.67	5.48	5.28	5.06	4.86	3.355
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.13	3.07	2.98	2.90	2.80	2.71	2.262
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.26	5.11	4.92	4.73	4.51	4.31	3.250
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.97	2.91	2.82	2.74	2.64	2.54	2.228
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.85	4.71	4.52	4.33	4.12	3.91	3.161

df														
11	2.201 / 3.106	2.40 / 3.60	2.50 / 3.80	2.61 / 4.02	2.70 / 4.21	2.79 / 4.40	2.86 / 4.54	2.95 / 4.74	3.01 / 4.88	3.09 / 5.07	3.20 / 5.32	3.36 / 5.67	3.59 / 6.22	3.98 / 7.20
														4.84 / 9.65
12	2.179 / 3.055	2.30 / 3.36	2.40 / 3.56	2.50 / 3.78	2.60 / 3.98	2.69 / 4.16	2.76 / 4.30	2.85 / 4.50	2.92 / 4.65	3.00 / 4.82	3.11 / 5.06	3.26 / 5.41	3.49 / 5.95	3.88 / 6.93
														4.75 / 9.33
13	2.160 / 3.012	2.21 / 3.16	2.32 / 3.37	2.42 / 3.59	2.51 / 3.78	2.60 / 3.96	2.67 / 4.10	2.77 / 4.30	2.84 / 4.44	2.92 / 4.62	3.02 / 4.86	3.18 / 5.20	3.41 / 5.74	3.80 / 6.70
														4.67 / 9.07
14	2.145 / 2.977	2.13 / 3.00	2.24 / 3.21	2.35 / 3.43	2.44 / 3.62	2.53 / 3.80	2.60 / 3.94	2.70 / 4.14	2.77 / 4.28	2.85 / 4.46	2.96 / 4.69	3.11 / 5.03	3.34 / 5.56	3.74 / 6.51
														4.60 / 8.86
15	2.131 / 2.947	2.07 / 2.87	2.18 / 3.07	2.29 / 3.29	2.39 / 3.48	2.48 / 3.67	2.55 / 3.80	2.64 / 4.00	2.70 / 4.14	2.79 / 4.32	2.90 / 4.56	3.06 / 4.89	3.29 / 5.42	3.68 / 6.36
														4.54 / 8.68
16	2.120 / 2.921	2.01 / 2.75	2.13 / 2.96	2.24 / 3.18	2.33 / 3.37	2.42 / 3.55	2.49 / 3.69	2.59 / 3.89	2.66 / 4.03	2.74 / 4.20	2.85 / 4.44	3.01 / 4.77	3.24 / 5.29	3.63 / 6.23
														4.49 / 8.53
17	2.110 / 2.898	1.96 / 2.65	2.08 / 2.86	2.19 / 3.08	2.29 / 3.27	2.38 / 3.45	2.45 / 3.59	2.55 / 3.79	2.62 / 3.93	2.70 / 4.10	2.81 / 4.34	2.96 / 4.67	3.20 / 5.18	3.59 / 6.11
														4.45 / 8.40
18	2.101 / 2.878	1.92 / 2.57	2.04 / 2.79	2.15 / 3.01	2.25 / 3.20	2.34 / 3.37	2.41 / 3.51	2.51 / 3.71	2.58 / 3.85	2.66 / 4.01	2.77 / 4.25	2.93 / 4.58	3.16 / 5.09	3.55 / 6.01
														4.41 / 8.28
19	2.093 / 2.861	1.88 / 2.49	2.00 / 2.70	2.11 / 2.92	2.21 / 3.12	2.31 / 3.30	2.38 / 3.43	2.48 / 3.63	2.55 / 3.77	2.63 / 3.94	2.74 / 4.17	2.90 / 4.50	3.13 / 5.01	3.52 / 5.93
														4.38 / 8.18
20	2.086 / 2.845	1.84 / 2.42	1.96 / 2.63	2.08 / 2.86	2.18 / 3.05	2.28 / 3.23	2.35 / 3.37	2.45 / 3.56	2.52 / 3.69	2.60 / 3.87	2.71 / 4.10	2.87 / 4.43	3.10 / 4.94	3.49 / 5.85
														4.35 / 8.10
21	2.080 / 2.831	1.81 / 2.36	1.93 / 2.58	2.05 / 2.80	2.15 / 2.99	2.25 / 3.17	2.32 / 3.31	2.42 / 3.51	2.49 / 3.65	2.57 / 3.81	2.68 / 4.04	2.84 / 4.37	3.07 / 4.87	3.47 / 5.78
														4.32 / 8.02
22	2.074 / 2.819	1.78 / 2.30	1.91 / 2.53	2.03 / 2.75	2.13 / 2.94	2.23 / 3.12	2.30 / 3.26	2.40 / 3.45	2.47 / 3.59	2.55 / 3.75	2.66 / 3.99	2.82 / 4.31	3.05 / 4.82	3.44 / 5.72
														4.30 / 7.94
23	2.069 / 2.807	1.76 / 2.26	1.88 / 2.48	2.00 / 2.70	2.11 / 2.89	2.20 / 3.07	2.28 / 3.21	2.38 / 3.41	2.45 / 3.55	2.53 / 3.71	2.64 / 3.94	2.80 / 4.26	3.03 / 4.76	3.42 / 5.66
														4.28 / 7.88

Degrees of Freedom for Smaller Mean Square

TABLE OF VALUES OF F FOR DESIGNATED DEGREES OF FREEDOM IN GREATER AND SMALLER MEAN SQUARE—*Continued*

Degrees of Freedom for Smaller Mean Square	Degrees of Freedom for Greater Mean Square														Values at t
	1	2	3	4	5	6	7	8	10	12	16	24	50	∞	
24	4.26 / 7.82	3.40 / 5.61	3.01 / 4.72	2.78 / 4.22	2.62 / 3.90	2.51 / 3.67	2.42 / 3.50	2.36 / 3.36	2.26 / 3.17	2.18 / 3.03	2.09 / 2.85	1.98 / 2.66	1.86 / 2.44	1.73 / 2.21	2.064 / 2.797
25	4.24 / 7.77	3.38 / 5.57	2.99 / 4.68	2.76 / 4.18	2.60 / 3.86	2.49 / 3.63	2.41 / 3.46	2.34 / 3.32	2.24 / 3.13	2.16 / 2.99	2.07 / 2.81	1.96 / 2.62	1.84 / 2.40	1.71 / 2.17	2.060 / 2.787
26	4.22 / 7.72	3.37 / 5.53	2.98 / 4.64	2.74 / 4.14	2.59 / 3.82	2.47 / 3.55	2.39 / 3.42	2.32 / 3.29	2.22 / 3.09	2.15 / 2.96	2.05 / 2.78	1.95 / 2.58	1.82 / 2.36	1.69 / 2.13	2.056 / 2.779
27	4.21 / 7.68	3.35 / 5.49	2.96 / 4.60	2.73 / 4.11	2.57 / 3.78	2.46 / 3.56	2.37 / 3.39	2.30 / 3.26	2.20 / 3.06	2.13 / 2.93	2.03 / 2.74	1.93 / 2.55	1.80 / 2.33	1.67 / 2.10	2.052 / 2.771
28	4.20 / 7.64	3.34 / 5.45	2.95 / 4.57	2.71 / 4.07	2.56 / 3.75	2.44 / 3.53	2.36 / 3.36	2.29 / 3.23	2.19 / 3.03	2.12 / 2.90	2.02 / 2.71	1.91 / 2.52	1.78 / 2.30	1.65 / 2.06	2.048 / 2.763
29	4.18 / 7.60	3.33 / 5.42	2.93 / 4.54	2.70 / 4.04	2.54 / 3.73	2.43 / 3.50	2.35 / 3.33	2.28 / 3.20	2.18 / 3.00	2.10 / 2.87	2.00 / 2.68	1.90 / 2.49	1.77 / 2.27	1.64 / 2.03	2.045 / 2.756
30	4.17 / 7.56	3.32 / 5.39	2.92 / 4.51	2.69 / 4.02	2.53 / 3.70	2.42 / 3.47	2.34 / 3.30	2.27 / 3.17	2.16 / 2.98	2.09 / 2.84	1.99 / 2.66	1.89 / 2.47	1.76 / 2.24	1.62 / 2.01	2.042 / 2.750
35	4.12 / 7.42	3.26 / 5.27	2.87 / 4.40	2.64 / 3.91	2.48 / 3.59	2.37 / 3.37	2.29 / 3.19	2.22 / 3.07	2.11 / 2.87	2.04 / 2.74	1.94 / 2.56	1.83 / 2.37	1.70 / 2.13	1.57 / 1.90	2.030 / 2.724
40	4.08 / 7.31	3.23 / 5.18	2.84 / 4.31	2.61 / 3.83	2.45 / 3.51	2.34 / 3.29	2.25 / 3.12	2.18 / 2.99	2.07 / 2.80	2.00 / 2.66	1.90 / 2.48	1.79 / 2.29	1.66 / 2.05	1.52 / 1.82	2.021 / 2.704
45	4.06 / 7.23	3.21 / 5.11	2.81 / 4.25	2.58 / 3.77	2.42 / 3.45	2.31 / 3.23	2.22 / 3.07	2.15 / 2.94	2.05 / 2.74	1.97 / 2.61	1.87 / 2.43	1.76 / 2.23	1.63 / 1.99	1.48 / 1.75	2.014 / 2.690
50	4.03 / 7.17	3.18 / 5.06	2.79 / 4.20	2.56 / 3.72	2.40 / 3.41	2.29 / 3.19	2.20 / 3.02	2.13 / 2.89	2.02 / 2.70	1.95 / 2.56	1.85 / 2.38	1.74 / 2.18	1.60 / 1.94	1.44 / 1.68	2.008 / 2.678

Degrees of Freedom for Smaller Mean Square														
60	2.000/2.660	1.39/1.60	1.56/1.87	1.70/2.12	1.81/2.32	1.92/2.50	1.99/2.63	2.10/2.82	2.17/2.95	2.25/3.12	2.37/3.34	2.52/3.65	2.76/4.13	3.15/4.98
70	1.994/2.648	1.35/1.53	1.53/1.82	1.67/2.07	1.79/2.28	1.89/2.45	1.97/2.59	2.07/2.78	2.14/2.91	2.23/3.07	2.35/3.29	2.50/3.60	2.74/4.07	3.13/4.92
80	1.990/2.638	1.32/1.49	1.51/1.78	1.65/2.03	1.77/2.24	1.88/2.42	1.95/2.55	2.06/2.74	2.12/2.87	2.21/3.04	2.33/3.26	2.49/3.56	2.72/4.04	3.11/4.88
90	1.987/2.632	1.30/1.45	1.49/1.75	1.64/2.00	1.76/2.21	1.86/2.39	1.93/2.53	2.04/2.72	2.11/2.84	2.20/3.01	2.32/3.23	2.47/3.53	2.71/4.01	3.10/4.85
100	1.984/2.626	1.28/1.43	1.48/1.73	1.63/1.98	1.75/2.19	1.85/2.37	1.92/2.51	2.03/2.69	2.10/2.82	2.19/2.99	2.30/3.21	2.46/3.51	2.70/3.98	3.09/4.82
125	1.979/2.616	1.25/1.37	1.45/1.69	1.60/1.94	1.72/2.15	1.83/2.33	1.90/2.47	2.01/2.66	2.08/2.79	2.17/2.95	2.29/3.17	2.44/3.47	2.68/3.94	3.07/4.78
150	1.976/2.609	1.22/1.33	1.44/1.66	1.59/1.92	1.71/2.13	1.82/2.31	1.89/2.45	2.00/2.63	2.07/2.76	2.16/2.92	2.27/3.14	2.43/3.45	2.66/3.91	3.06/4.75
200	1.972/2.601	1.19/1.28	1.42/1.62	1.57/1.88	1.69/2.09	1.80/2.28	1.87/2.41	1.98/2.60	2.05/2.73	2.14/2.89	2.26/3.11	2.42/3.41	2.65/3.88	3.04/4.71
300	1.968/2.592	1.15/1.22	1.39/1.59	1.55/1.85	1.68/2.06	1.79/2.24	1.86/2.38	1.97/2.57	2.04/2.70	2.12/2.86	2.25/3.08	2.41/3.38	2.64/3.85	3.03/4.68
400	1.966/2.588	1.13/1.19	1.38/1.57	1.54/1.84	1.67/2.04	1.78/2.23	1.85/2.37	1.96/2.56	2.03/2.69	2.12/2.85	2.24/3.06	2.40/3.37	2.63/3.83	3.02/4.66
500	1.965/2.586	1.11/1.16	1.38/1.56	1.54/1.83	1.66/2.03	1.77/2.22	1.85/2.36	1.96/2.55	2.03/2.68	2.11/2.84	2.23/3.05	2.39/3.36	2.62/3.82	3.01/4.65
1000	1.962/2.581	1.08/1.11	1.36/1.54	1.53/1.81	1.65/2.01	1.76/2.20	1.84/2.34	1.95/2.53	2.02/2.66	2.10/2.82	2.22/3.04	2.38/3.34	2.61/3.80	3.00/4.63
∞	1.960/2.576	1.00/1.00	1.35/1.52	1.52/1.78	1.64/1.99	1.75/2.18	1.83/2.32	1.94/2.51	2.01/2.64	2.09/2.80	2.21/3.02	2.37/3.32	2.60/3.78	2.99/4.60

Degrees of Freedom for Smaller Mean Square	
60	4.00/7.08
70	3.98/7.01
80	3.96/6.96
90	3.95/6.92
100	3.94/6.90
125	3.92/6.84
150	3.90/6.81
200	3.89/6.76
300	3.87/6.72
400	3.86/6.70
500	3.86/6.69
1000	3.85/6.66
∞	3.84/6.64

BIBLIOGRAPHY

ADAMS, THOMAS S. (ed.), *Manual of Charting*, New York, Prentice-Hall, 1924.

ALLEN, R. G. D., *Mathematical Analysis for Economists*, New York, The Macmillan Company, 1939, 458 pages.

ARKIN, HERBERT, and COLTON, RAYMOND R., *Graphs*, New York and London, Harper & Brothers, 1936, 220 pages.

BABSON, ROGER W., *Business Barometers for Anticipating Conditions*, 21st ed., Babson Park, Babson's Statistical Organization Incorporated, 1929, 373 pages.

BAKER, O. E.; BORSODI, RALPH; and WILSON, M. L., *Agriculture in Modern Life*, New York and London, Harper & Brothers, 1939, 297 pages.

BANISTER, H., *Elementary Applications of Statistical Method*, London and Glasgow, Blackie and Son Limited, 1929, 56 pages.

Barlow's Tables of Square Cubes, Square R ots, Cube Roots, and Reciprocals, edited by L. J. COMRIE, New York, Spon and Chamberlain, 1930.

BATEN, WILLIAM DOWELL, *Elementary Mathematical Statistics*, New York, John Wiley & Sons, 1938, 338 pages.

BERNHEIM, ALFRED L. (ed.), *Big Business: Its Growth and Its Place*, New York, Twentieth Century Fund, 1937, 102 pages.

BERRIDGE, WM. A., *Cycles of Unemployment*, Boston, Houghton Mifflin Company, 1923.

BEVERIDGE, W. H., *Unemployment: A Problem of Industry*, New York, Longmans, Green & Company, 1930, 541 pages.

BLACKETT, O. W., and WILSON, W. P., "A Method of Isolating Sinusoidal Components in Economic Time Series," *Michigan Business Studies*, University of Michigan, 1938, 390 pages.

BOWLEY, ARTHUR L., *Elements of Statistics*, 5th ed., New York, Charles Scribner's Sons, 1926 (one-volume edition), 462 pages.

BROWN, T. H., *Problems in Business Statistics*, New York, McGraw-Hill Book Company, 1931.

BURGESS, ROBERT W., *Introduction to the Mathematics of Statistics*, Boston, Houghton Mifflin Company, 1927.

BURNS, ARTHUR F., *Production Trends in the United States Since 1870*, New York, National Bureau of Economic Research, 1934, 353 pages.

BUROS, OSCAR KRISEN (ed.), *Research and Statistical Methodology Books and Reviews of 1933–1938*, New Brunswick, Rutgers University Press, 1938, 100 pages.

CHADDOCK, R. E., *Principles and Methods of Statistics*, Boston, Houghton Mifflin Company, 1925.

CHADDOCK, R. E., and CROXTON, F. E., *Exercises in Statistical Methods*, Boston, Houghton Mifflin Company, 1928.

CONNOR, L. R., *Statistics in Theory and Practice*, New York, Isaac Pitman and Sons, 1933.

COWAN, DONALD, R. G., *Sales Analysis from the Management Standpoint*, Chicago University of Chicago Press, 1938, 210 pages.

COWLES, ALFRED, 3RD, *Common-Stock Indexes 1871–1937*, Bloomington, Indiana, Principia Press, 1938, 499 pages.

CRUM, W. L., and PATTON, A. C., *Economic Statistics*, New York, A. W. Shaw Company, 1928.

CROXTON, FREDERICK E., and COWDEN, DUDLEY J., *Applied General Statistics*, New York, Prentice-Hall, 1939, 944 pages.

DAVIES, G. R., and CROWDER, W. F., *Methods of Statistical Analysis*, New York, John Wiley & Sons, 1933, 355 pages.

DAVIS, H. T., and NELSON, W. F. C., *Elements of Statistics*, Bloomington, The Principia Press, 2d ed., 1937.

DAY, EDMUND E., *Statistical Analysis*, New York, The Macmillan Company, 1925.

DITTMER, CLARENCE G., *Introduction to Social Statistics,* New York, A. W. Shaw Company, 1926.

DUBLIN, LOUIS I., *Population Problems*, Boston, Houghton Mifflin Company, 1926, 318 pages.

DUNLAP, J. W., and KURTZ, A. K., *Handbook of Statistical Nomographs, Tables, and Formulas*, Yonkers-on-Hudson, World Book Company, 1932.

EDIE, LIONEL, D. (ed.), *The Stabilization of Business*, New York, The Macmillan Company, 1923, 400 pages.

EVANS, G. C., *Mathematical Introduction to Economics*, New York, McGraw-Hill Book Company, 1930.

EZEKIEL, MORDECAI, *Methods of Correlation Analysis*, New York, John Wiley & Sons, 1930.

FISHER, A., *An Elementary Treatise on Frequency Curves*, New York, The Macmillan Company, 1923, 240 pages.

—— *The Mathematical Theory of Probabilities and Its Application to Frequency Curves and Statistical Methods*, New York, The Macmillan Company, 1922, 289 pages.

FISHER, IRVING, *Mathematical Investigation in the Theory of Value and Prices*, New Haven, Yale University Press, 1926.

FISHER, R. A., *Statistical Methods for Research Workers*, London, Oliver and Boyd, 4th ed., 1932.

FISHER, R. A., and YATES, F., *Statistical Tables for Biological, Agricultural and Medical Research*, London, Oliver and Boyd, 1938, 90 pages.

FLORENCE, P. SARGENT, *The Statistical Method in Economics and Political Science*, New York, Harcourt, Brace & Company, 1929.

GAY, EDWIN F., MITCHELL, WESLEY C., and others, *Recent Economic Changes*, New York, National Bureau of Economic Research, 1929, 950 pages (2 volumes).

GLOVER, J. W., *Tables of Applied Mathematics, Finance, Insurance, Statistics*, 2d ed., Ann Arbor, Mich., George Wahr, 1923, 678 pages.

HABERLER, GOTTFRIED, *Der Sinn der Indexzahlen*, Tübingen, J. C. B. Mohr, 1927.

HANEY, LEWIS H., *Business Forecasting*, Boston, Ginn & Company, 1931, 373 pages.

HANSEN, A. H., *Business Cycle Theory*, Boston, Ginn & Company, 1927.

HARDY, CHARLES, O., *Risk and Risk Bearing*, The University of Chicago Press, 1923, 400 pages.

HARDY, CHARLES O., and COX, GARFIELD, V., *Forecasting Business Conditions* New York. The Macmillan Company, 1927.

HARPER, F. H., *Elements of Practical Statistics*, New York, The Macmillan Company, 1930.

HASKEL, ALLAN C., *Graphic Charts in Business*, New York, Codex Book Company, 1922.

HODGMAN, C. D., *Mathematical Tables* (from "Handbook of Chemistry and Physics"), Cleveland, Chemical Rubber Publishing Company, 1931.

JEROME, HARRY, *Migration and Business Cycles*, New York, National Bureau of Economic Research, Inc., 1926, 256 pages.

—— *Statistical Method*, New York, Harper & Brothers, 1924.

JORDAN, DAVID F., *Practical Business Forecasting*, New York, Prentice-Hall, 1927.

Journal of the American Statistical Association, published quarterly by the American Statistical Association, New York.

KARSTEN, K. G., *Charts and Graphs*, New York, Prentice-Hall, 1925.

KELLEY, TRUMAN, L., *Statistical Method*, New York, The Macmillan Company, 1923. Also cf. *The Kelley Statistical Tables*.

KENNEY, JOHN F., *Mathematics of Statistics*, Part One and Part Two, New York, D. Van Nostrand Company, 1939, 248 and 202 pages.

KING, WILLFORD I., *Employment, Hours, and Earnings in Prosperity and Depression*, New York, National Bureau of Economic Research, 1923, 147 pages.

—— *Index Numbers Elucidated*, New York, Longmans, Green & Company, 1930.

—— *The National Income and Its Purchasing Power*, New York, National Bureau of Economic Research, Inc., 1930.

KURTZ, ALBERT K., and EDGERTON, HAROLD A., *Statistical Dictionary of Terms and Symbols*, New York, John Wiley & Sons, 1939, 191 pages.

KUZNETS, SIMON, *Commodity Flow and Capital Formation*, Vol. I, New York, National Bureau of Economic Research, 1938, 505 pages.

—— *National Income and Capital Formation 1919–1935*, A Preliminary Report, New York, National Bureau of Economic Research, 1937, 86 pages.

—— *Secular Movements in Production and Prices*, Boston, Houghton Mifflin Company, 1930.

—— *Seasonal Variations in Industry and Trade*, New York, National Bureau of Economic Research, 1933.

LACROIX and RAGOT, *A Graphic Table Combining Logarithms and Antilogarithms*, New York, The Macmillan Company, 1927.

LOVITT, W. V., and HOLTZCLAW, H. F., *Statistics*, New York, Prentice-Hall, 1929.

MACAULAY, F. R., *The Smoothing of Time Series*, New York, National Bureau of Economic Research, 1931.

MCMILLAN, A. W., *Measurement in Social Work*, Chicago, University of Chicago Press, 1930.

MILLS, FREDERICK C., *Statistical Methods Applied to Economics and Business*, revised, New York, Henry Holt & Company, 1938, 746 pages.

—— *The Behavior of Prices*, New York, National Bureau of Economic Research, 1927.

MILLS, F. C., and DAVENPORT, D. H., *A Manual of Problems and Tables in Statistics with Notes on Statistical Procedure*, New York, Henry Holt & Company, 1925, 203 pages.

MISES, RICHARD VON, *Probability, Statistics and Truth*, New York, The Macmillan Company, 1939, 323 pages.

MITCHELL, WESLEY C., *Business Cycles: The Problem and Its Setting*, New York, National Bureau of Economic Research, 1927.

MOORE, HENRY LUDWELL, *Economic Cycles: Their Law and Cause*, New York, The Macmillan Company, 1914, 149 pages.

—— *Generating Economic Cycles*, New York, The Macmillan Company, 1923, 141 pages.

—— *Synthetic Economics*, New York, The Macmillan Company, 1929.

MOULTON, HAROLD G., *The Formation of Capital*, Washington, D. C., The Brookings Institution, 1935, 201 pages.

—— *Income and Economic Progress*, Washington, D. C., The Brookings Institution, 1935, 188 pages.

MUDGETT, BRUCE D., *Statistical Tables and Graphs*, Boston, Houghton Mifflin Company, 1930.

NAGEL, ERNEST, *Principles of the Theory of Probability*, Chicago, University of Chicago Press, 1939, 80 pages.

NOURSE, EDWIN G., and ASSOCIATES, *America's Capacity to Produce*, Washington, D. C., The Brookings Institution, 1934, 597 pages.

OLIVIER, MAURICE, *Les nombres indices de la variation des prix*, Paris, Marcel Giard, 1927.

PASSANO, L. M., *Calculus and Graphs*, New York, The Macmillan Company, 1932.

PATON, W. A., *Corporate Profits as Shown by Audit Reports*, New York, National Bureau of Economic Research, 1935, 151 pages.

PEARSON, KARL, *Tables for Statisticians and Biometricians*, Cambridge University Press, 1914, 114 pages.   (Out of print.)

PEIRCE, B. O., *A Short Table of Integrals*, Boston, Ginn & Company, 1910.

PERSONS, W. M., *The Construction of Index Numbers*, Boston, Houghton Mifflin Company, 1928.

—— *Forecasting Business Cycles*, New York, John Wiley & Sons, 1931.

PHILIP, M., *Principles of Financial and Statistical Mathematics*, New York, Prentice-Hall, 1932.

Recent Economic Changes, Report of the Committee on Recent Economic Changes of the President's Conference on Unemployment, New York, McGraw-Hill Book Company, 1929.   Two volumes.

REINHARDT, J. M., and DAVIES, G. R., *Principles and Methods of Sociology*, New York, Prentice-Hall, 1932.

RICE, S. A., *Methods in Social Science*, Chicago, University of Chicago Press, 1931.

—— (Ed.), *Statistics in Social Studies*, Philadelphia, University of Pennsylvania Press, 1930.

RIDER, PAUL R., *An Introduction to Modern Statistical Methods*, New York, John Wiley & Sons, 1939, 220 pages.

RIEGEL, ROBERT, *Elements of Business Statistics*, rev. ed., New York, D. Appleton & Company, 1927, 549 pages.

RIETZ, H. L. (ed.), *Handbook of Mathematical Statistics*, Boston, Houghton Mifflin Company, 1924.

RIGGLEMAN, J. R., and FRISBEE, I. N., *Business Statistics*, revised ed., New York, McGraw-Hill Book Company, 1938, 790 pages.

SCHLUTER, W. C., *How To Do Research Work*, New York, Prentice-Hall, 1927, 137 pages.

SCHULTZ, HENRY, *Statistical Laws of Demand and Supply*, Chicago, University of Chicago Press, 1928.

—— *The Theory and Measurement of Demand*, Chicago, The University of Chicago Press, 1938, 817 pages.

SCHUMPETER, J. A., *Business Cycles*, Vols. I and II, New York, McGraw-Hill Book Company, 1939, 1094 pages.

SECRIST, HORACE, *An Introduction to Statistical Methods*, New York, The Macmillan Company, 1925, 584 pages.

—— *Banking Ratios*, Claremont Colleges Research Studies I, Stanford University Press, 1930, 608 pages.

SMITH, JOHN H., *Tests of Significance: What They Mean and How To Use Them*, Chicago, University of Chicago Press, 1939, 90 pages.

SNEDECOR, GEORGE W., *Statistical Methods Applied to Experiments in Agriculture and Biology*, Ames, Iowa, Collegiate Press, 1937, 341 pages.

SNIDER, JOSEPH L., *Business Statistics*, 2d ed., New York, McGraw-Hill Book Company, Inc., 1932.

SNYDER, CARL, *Business Cycles and Business Measurements*, New York, The Macmillan Company, 1927.

STEWART, PAUL W., and DEWHURST, J. FREDERIC, *Does Distribution Cost Too Much?*, New York, The Twentieth Century Fund, 1939, 403 pages.

STOCKTON, JOHN R., *An Introduction to Business Statistics*, Boston, D. C. Heath & Company, 1938, 378 pages.

STOCKWELL, H. G., *How to Read a Financial Statement*, New York, The Ronald Press Company, 1925, 443 pages.

—— *How to Read a Profit and Loss Statement*, New York, The Ronald Press Company, 1927, 411 pages.

Survey of Current Business, Published monthly by the United States Department of Commerce, Washington, D. C.

SUTCLIFFE, WILLIAM G., *Elementary Statistical Methods*, New York, McGraw-Hill Book Company, 1925, 338 pages.

—— *Statistics for the Business Man*, New York, Harper & Brothers, 1930, 243 pages.

THOMAS, DOROTHY S., *Social Aspects of the Business Cycle*, New York, A. A. Knopf, 1927.

THURSTONE, L. L., *Fundamentals of Statistics*, New York, The Macmillan Company, 1925, 237 pages.

TINBERGEN, J., *A Method and Its Application to Investment Activity*, Geneva, League of Nations Economic Intelligence Service, 1939, 164 pages.

—— *Business Cycles in the United States of America 1919–1932*, Geneva, League of Nations Economic Intelligence Service, 1939, 244 pages.

TINTNER, GERHARD, *The Variate Difference Method*, Indiana, Principia Press, 1940, 167 pages.

TRELOAR, ALAN E., *Elements of Statistical Reasoning*, New York, John Wiley & Sons, 1939, 261 pages.

VANDERBLUE, HORACE N., *Problems in Business Economics*, New York, A. W. Shaw Company, 1929.

WAGEMANN, ERNST, *Economic Rhythm*, New York, McGraw-Hill Book Company, Inc., 1930.

WALKER, HELEN M., *Studies in the History of Statistical Method*, Baltimore, Williams and Wilkins Company, 1929, 229 pages.

WALSH, C. M., *The Measurement of General Exchange Value*, New York, The Macmillan Company, 1901.

WAUGH, ALBERT E., *Elements of Statistical Method*, New York. McGraw-Hill Book Company, 1938, 381 pages.

WHITE, R. CLYDE, *Social Statistics*, New York, Harper and Brothers, 1933, 471 pages.

WOLFENDEN, H. H., *Population Statistics*, New York, Actuarial Society of America, 1925.

World Economic Survey, Eighth Year, 1938–1939, Geneva, League of Nations, 1939, 247 pages.

YOUNG, BENJAMIN F., *Statistics as Applied in Business*, New York, The Ronald Press Company, 1925, 639 pages.

YULE, G. UNDY, and KENDALL, M. G., *An Introduction to the Theory of Statistics*, London, Chas. Griffin and Company, 1937, 570 pages.

CLASSIFIED READINGS FROM READILY AVAILABLE TEXTS

AUTHOR	TITLE	CODE LETTER
BOWLEY, A. L.,	*Elements of Statistics*	A
BURGESS, R. W.,	*Mathematics of Statistics*	B
CHADDOCK, R. E.,	*Principles and Methods of Statistics*	C
CROXTON AND COWDEN,	*Practical Business Statistics*	D
CROXTON and COWDEN,	*Applied General Statistics*	DD
DAVIES AND CROWDER,	*Methods of Statistical Analysis*	E
DAY, E. E.,	*Statistical Analysis*	F
FISHER, IRVING,	*Making of Index Numbers*	G
JEROME, HARRY,	*Statistical Method*	H
MILLS, F. C.,	*Statistical Methods* (Revised)	J
MUDGETT, B. D.,	*Statistical Tables and Graphs*	K
RICHARDSON, C. H.,	*Introduction to Statistical Analysis*	L
RIGGLEMAN AND FRISBEE,	*Business Statistics* (2d ed.)	M
SECRIST, HORACE,	*Introduction to Statistical Methods*	N
SMITH, JAMES G.,	*Elementary Statistics*	O
STOCKTON, JOHN R.,	*Introduction to Business Statistics*	P
WAUGH, ALBERT E.,	*Elements of Statistical Method*	Q

1. Introduction
   A: 3–17
   B: 1–15
   C: 3–39
   D: Chapter I
   DD: 1–14

   E: 1–3
   F: 1–9, 25–35
   J: 1–7
   L: 1–14
   M: 1–13

   N: 1–20
   O: 3–21
   P: 1–15
   Q: 1–12

2. Collection of data
   A: 18–51
   C: 10, 371–382
   D: Chapter II
   DD: 15–37

   E: 4–8
   F: 383–389
   M: 14–47
   N: 47–70

   O: 22–37
   P: 16–34
   Q: 354–364

3. Readily available data
   D & DD: Appendix A
   F: 389–393

   M: 36–47
   N: 22–46

   O: 72–93

4. Classification
   C: 43–51
   F: 36–48

   K: 3–15
   N: 124–128

   P: 35–43

5. Tabulation
   A: 52–81
   C: 398–403
   D: 34–37
   DD: 37–44

   E: 8
   F: 393–403
   K: 16–28
   M: 48–75

   N: 128–138
   P: 44–56
   Q: 339–341

6. Table structure

C:	403–411	K:	29–60	O:	38–42
D:	Chapter III	M:	63–70	P:	57–72
DD:	49–68	N:	138–170	Q:	339–342
F:	403–411				

7. Graphics—bar charts

C:	418–420, 429–432	F:	404–420	N:	171–200
D:	110–121	J:	41–42, 63–75	P:	74–76
DD:	124–130	K:	60–89	Q:	343
E:	13–14	M:	83–90		

8. Graphics—time charts

C:	428–429	E:	20–25	N:	242–259
D:	Chapter IV	K:	122–145	P:	76–78, 145–148
DD:	70–119	M:	96–108	Q:	343–345

9. Graphics—ratio charts

C:	433–438	J:	26–41	O:	63–71
D & DD:	Chapter V	K:	145–160	P:	148–156
E:	20–25	M:	109–122	Q:	346–353
F:	246–257	N:	243–255		

10. Graphics—statistical maps

C:	432–433	F:	211–231	N:	201–206
D:	121–129	K:	160–192	O:	50–54
DD:	137–145	M:	152–155	P:	82–84

11. Graphic rules

C:	439–445	F:	411–420	M:	122–125
DD:	79–91	J:	32–35	Q:	342–346

12. Rates and ratios

B:	16–40	D:	Chapter VII	DD:	146–163

13. Logarithms, squares, etc.

D:	Appendices C and H	E:	331–344	M:	723–759
DD:	Appendices O and P	J:	22–26	P:	Appendices G and H

14. Graphs of equations

C:	427–428	J:	8–32	O:	94–95
D:	65–73	K:	90–101		

15. Frequency distributions

B:	53–60	E:	8–32	P:	51–56
DD:	164–189	K:	102–120	Q:	13–21

  (a) Arrays, frequency tables, etc.

C:	53–61	F:	62–75	L:	15–22
D:	151–157	J:	50–63	M:	132–141

  (b) Graphs of frequency distributions

C:	61–79	J:	63–80	N:	221–232
D:	157–168	L:	22–28	O:	54–60
F:	75–80	M:	141–146		

(c) Cumulative frequency distributions

B:	61–69	F:	80–89	M:	146–149
C:	115–122	J:	80–85	N:	232–242
D:	165–168	L:	33–36		

16. Introduction to averages

A:	82	J:	86–101	O:	105–108
B:	80–81	M:	164–166	P:	92–95
C:	81–87	N:	261–264	Q:	22
F:	134–135				

17. Arithmetic mean

A:	82–94	E:	33–38	N:	264–282
B:	81–85	F:	135–139	O:	108–110
C:	87–105	J:	101–109	P:	96–102
D:	169–176, 183–193	L:	43–53	Q:	22–24, 36–42
DD:	194–207	M:	166–170		

18. Geometric mean

A:	107–108	E:	38–41	N:	307–309
B:	96–98	F:	140–141	O:	112
C:	125–127	J:	125–132	P:	103–105
D:	178–182	L:	63–67	Q:	27–29
DD:	221–226	M:	176–178		

19. Harmonic mean

B:	90–94	E:	41–43	L:	67–68
C:	p. 87 note	F:	141–142	O:	113
D:	182–183	J:	132–133	Q:	29–30
DD:	226–232				

20. Median, modes, and quartiles

A:	96–107	DD:	207–221	M:	170–176
B:	85–89	E:	43–48	N:	282–303
C:	{ 107–115, 122–124 129–134, 139–147	F:	142–151	O:	110–112, 137–40
		J:	109–125	P:	105–112
D:	176–178, 193–203	L:	53–62	Q:	24–27, 30–33, 42–53

21. Comparison of averages

A:	108–109	E:	49–53	M:	164–166
B:	94–96, 98–107	F:	151–162	N:	310–322
C:	134–139, 148–149	J:	134–136	P:	112–118
D:	201–203	L:	68–76	Q:	53–57
DD:	228–232				

22. Introduction to dispersion

A:	110–111	F:	163–165	N:	324–335
B:	135–136	J:	137–139	P:	119–122
C:	150–153	M:	215–217	Q:	59–63
DD:	234–237				

AUTHOR INDEX

SUBJECT INDEX